OUTLOOK
Alzheimer's disease

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE



MALARIA BITES BACK

The race to stop a killer in southeast Asia before drug resistance spreads

PAGE 458

ATOMIC PHYSICS

PHOTON-FREE QUANTUM OPTICS

Tunable matter waves from ultracold atoms

PAGES 481 & 589

NEUROBIOLOGY

ATTRACTIVE PROSPECT

The neural circuit that drives mate choice in fruit flies

PAGES 485 & 564

INSIGHT

THE TROPICS

Challenges in disease control and biodiversity

PAGE 489

NATURE.COM/NATURE
26 July 2018
Vol. 559, No. 7715

THIS WEEK

EDITORIALS

CONSERVATION How to publish data and still protect rare species **p.444**

WORLD VIEW Preprint boom could worsen media coverage of science **p.445**



HISS-STORY World's oldest snake found sealed inside amber **p.447**

Pay people to fight malaria

Workers on the front line of the battle against the disease are too often overlooked as scientists and funders plot how to defeat malaria.

Nature sent a reporter to Myanmar, Cambodia and Thailand in May to see how researchers are helping countries in the Greater Mekong region reach their goal of eliminating malaria by 2030. Her News Feature on page 458 highlights how political, financial and geographical obstacles allow the disease to persist. The broader malaria community, including researchers and funders, must do more to address these challenges. It is a race against time. The disease must be stamped out in the region now, while it remains relatively rare. Allow it to increase, and the chance will be lost. Worse, drug-resistant parasites there will multiply and spread.

It's no secret that the best way to beat malaria is to ensure that every infected person can quickly reach a health worker and be given artemisinin-based combination therapies (ACTs) — pills that swiftly kill the parasites that cause malaria before they spread to other people. The importance of health workers is highlighted by what happens in their absence. Hundreds of thousands of African children die every year because they do not get the pills in time. And in Venezuela — once declared malaria-free — the disease has come back with a vengeance as political instability has fractured health services.

A review of 75 malaria resurgences around the world from 1930 to 2011 found that 90% were caused by lapses in basic malaria-control measures — mosquito control (mainly through insecticides) and health systems (J. M. Cohen *et al. Malaria J.* **11**, 122; 2012). Insecticides are potent tools for reducing malaria, but resistance is on the rise. Bed nets also reduce risk, but can't always prevent the disease. A strong public-health system is necessary because there is no powerful vaccine for malaria, unlike other illnesses subject to elimination campaigns, such as polio.

Yet health systems in many malaria-endemic nations are riddled with holes. Malaria health workers are often unsalaried, so many leave for other jobs. Networks of researchers, technicians and administrators support health workers, provide surveillance and inform policy. But health budgets in low-income nations are often too small to attract enough of these skilled workers.

At present, too many of these crucial positions are unpaid. The Bill & Melinda Gates Foundation in Seattle, Washington, the US government and many other donors do not pay salaries for health workers and other staff in national health systems. The Global Fund to Fight AIDS, Tuberculosis and Malaria is a rare exception in supplementing pay for national health workers — but even it does not allocate funds to researchers and officials at higher levels in ministries of health. It says that national governments should do this, but many can't or don't.

The problem is complex, but there are a few ways in which the scientific community could help. Researchers can analyse the costs and value of health-system components and compare them with those of other projects, such as genetically modifying mosquitoes to wipe them out, or conducting vaccine studies. This would help donors and governments re-evaluate their priorities when funds are limited. Researchers can also help by developing simple tools and treatments for other diseases, thus allowing local malaria health workers to be more versatile. In this way, their services remain crucial even as malaria infections drop — as they have across southeast Asia. This added value means that governments might be more willing to pay workers.

Furthermore, top-tier scientists in Western countries can work to boost the status of their colleagues in malaria-endemic regions. After all, local researchers working in government agencies are best placed to lobby their leaders for bigger health budgets. At the moment, these local researchers often provide a supporting role to those from the United States and Europe who are the lead investigators on large grants. The international scientific community should put more emphasis on providing long-term mentorship, building laboratories and helping colleagues to boost their careers.

"The best way to beat malaria is to ensure that every infected person can quickly reach a health worker."

And — despite strong disincentives for doing so — experts should push back against funders and journals that favour 'outside the box' ideas over routine support for health workers in remote villages. For example, drones have been touted as a new means of delivering malaria drugs to remote areas. Researchers should ask for evidence that such expenditures are more valuable than improving the basic health system.

Last month, the World Health Organization declared Paraguay to be malaria-free, even as cases were rising in Venezuela, Panama, Peru and Nicaragua. Public-health officials attributed the nation's victory to its health system's ability to quickly detect, treat and track cases. Sometimes, the classic recipes are the best. ■

Inspiring women

Awards highlight impressive efforts that reach the next generation.

In April, *Nature* announced two annual prizes: one recognizes excellent early-career female scientists, and the other acknowledges individuals or organizations that have led grass-roots initiatives to support increased access to, or interest in, science, technology, engineering and mathematics (STEM) for girls and young women around the world. This week, we announce the candidates who have made the longlist of ten for each Nature Research award, and congratulate all of them. The impressive breadth and quality of all the entries deserve wide recognition, too.

There are many reasons for which women are discouraged from entering STEM professions or from staying in them. They include

© 2018 Springer Nature Limited. All rights reserved.

unequal access to education; conscious and unconscious discrimination in promotion, funding, peer review, citation and more; unequal pay; and, often, a disproportionate burden of domestic responsibilities. All of that can help to explain why women still make up only around 30% of professional researchers. It is encouraging to see the many entries from women who are successful researchers and who are also determined to help others and to inspire the next generation. One route towards equality in science is to offer positive role models to young people.

Some of the most impressive prize entrants focus on countries and locations in the global south, working with particularly impoverished and underprivileged communities. This often comes with additional challenges — sometimes, just reaching remote locations to mentor girls and encourage them to pursue science requires serious commitment. For example, one project, Cielo y Tierra, organizes outreach efforts in rural South America and Africa.

Obviously, not all of the girls and young women reached in this way will end up pursuing science in their studies or careers. But awakening their curiosity and nurturing their ambitions is still worthwhile.

It turns out that mentorship and support over long distances can be highly effective. One scheme has successfully matched researchers with young girls from low-income communities as penfriends. Working with US schools and with scientists around the world, 'Letters to a Pre-Scientist' fosters a love for science, offers support and provides personal experiences. Scientists are encouraged to write at least once every few months, and to give instances of how science has, for example, taught them to persevere or to learn from failure. It's a simple and effective way to demystify science by creating personal connections. Perhaps you should try it.

For students already in the university environment, gaining access to female scientists might still be difficult. Therefore, many award applicants have set up visiting female researchers' programmes, lecture series focused on female speakers, or social events that enable networking with prominent female researchers.

"One route towards equality in science is to offer positive role models to young people."

In the global north, many of the efforts focus on girls and young women from minority ethnic groups. Others target communities of refugees, the children of female prison inmates and those who live in orphanages.

Many of these projects, the prize entrants told us, are driven by the entrants' own experiences as women in science. For example, some scientists who are also mothers are working to improve childcare provision at scientific conferences. Others have overcome particularly strong local cultural constraints, under which women are rarely seen in public, let alone in science and education.

All of the entrants believe that women should have the opportunity, agency and support to become researchers. We agree. (And we appreciate that gender is neither binary nor necessarily fixed.) Together, these examples show the enormous range of ways to inspire girls and young women, and they offer lessons for all. The full longlists for the two prizes — the Inspiring Science Award and the Innovating Science Award — are available at go.nature.com/2jmri9z. The winners will be announced in October. ■

Publish not perish

Conservationists have a new tool for weighing up the risks and benefits of open data.

From 2013 to 2017, listening to the gentle 'ding ding' of the night parrot was forbidden. Long feared extinct, when the bird (*Pezoporus occidentalis*) was rediscovered, officials in Australia decided to restrict both location data and recordings of its signature call for fear that poachers and enthusiasts might use the information to track and disturb the creatures. Yet when the recordings were declassified and shared last year, conservationists were delighted by what followed: at least three new populations have since been discovered by people using the call to recognize the birds.

Since the bird's rediscovery in 2013, the Australian government has put in place proper conservation safeguards, such as making it illegal to approach the creature's habitat. It's a good example of authorities weighing up the risks and benefits of publicizing the discovery of a rare species and then reaching a sensible compromise. That kind of decision process is increasingly in demand, as data sources and sharing proliferate beyond conventional academic audiences in ways that risk, for example, helping hunters and illegal wildlife traders to track down target species.

In a Perspective published this week in *Nature Ecology & Evolution* (A. I. T. Tulloch *et al. Nature Ecol. Evol.* **2**, 1209–1217; 2018), conservation experts offer a way to help scientists and officials to decide when to publish such sensitive information — and when not to. It's the latest development in an ongoing debate that pits advocates of open data against those who take a harder line and want more restrictions. The authors warn that a default position in which location data are withheld if a species is identified as being of high biological significance and under high threat — as recommended by the Global Biodiversity Information Facility — risks missing out on the benefits of data sharing.

To aim for a more balanced approach, the scientists drew up a decision tree to help people judge what to do with information gained from wildlife monitoring and surveys. A series of steps asks questions such as "Could data be used to mitigate threats to species?" and "Would sharing location data increase risk of species decline through increased visitation?" In some cases — fish spawning locations for one, because the fishing industry would love to target them — the recommendation is to keep everything from the name of a species to its location under wraps. But in other cases, the need for secrecy is trumped by the possible benefits of transparency. Open data could help local communities fight to protect a habitat when development is threatening a species.

Ayesha Tulloch, an environmental scientist at the University of Sydney in Australia who led the analysis, says her team was surprised by the low number of examples they tried that produced a 'don't publish' decision. That, she adds, could help scientists to get funding for projects that might otherwise be rejected because of the restrictions expected to be placed on the eventual results. Several government departments in Australia and New Zealand — including those who have historically kept data from the public — have already said that they are interested in using the decision framework to help set policy.

Such discussions are timely. Scientists are poised to launch a satellite-tracking project that will massively increase the number of species worldwide whose locations can be traced in near-real time. Earlier this year, a satellite antenna called ICARUS — International Cooperation for Animal Research Using Space — was attached to the International Space Station. Combined with lightweight animal tags, it will allow researchers to follow the long-range movements of much smaller species than has previously been possible. For some species, that information could place them at greater risk. There are already concerns that unethical safari guides and hunters have hacked into the biotagging systems used to keep tabs on endangered animals such as lions (S. J. Cooke *et al. Conserv. Biol.* **31**, 1205–1207; 2017).

About one in six of the threatened species listed by the International Union for Conservation of Nature are classed as data-deficient, which means conservationists and scientists face a struggle to perform basic analyses such as population modelling and to work out which habitats must be preserved. Data bring power. But they also bring responsibility. This new decision-making tool could help to generate the former while respecting the latter. ■

© 2018 Springer Nature Limited. All rights reserved.



Preprints could promote confusion and distortion

The scientific community must take measures to keep preprints from distorting the public's understanding of science, says **Tom Sheldon**.

Thousands of papers are submitted every month to the platforms arXiv and bioRxiv, which make manuscripts available before they have been peer reviewed and accepted by a journal. Scientists applaud preprints because they enable researchers to claim priority and make their findings available more quickly, unshackled from sluggish and tyrannical journals.

This might make sense within the scientific community, but as someone who has worked for years with researchers and journalists to ensure responsible coverage of science in the media, I fear that this method of publication holds substantial risks for the broader community — risks that are not being given proper consideration by the champions of preprint. Weak work that hasn't been reviewed could get overblown in the media. Conversely, better work could be ignored.

Those who run preprint servers, researchers who submit to them and anyone considering publicizing a preprint have a responsibility to take note of these risks. The responses we received to an open letter we posted on 17 July (see go.nature.com/2uxf86x) have not entirely reassured me that this is the case. Here, I call on scientists to reflect on — and find ways to limit — the potential of preprints to elicit public confusion and misinformation.

Many people still learn about science the same way they learn about Vladimir Putin, Syria or the World Cup: through news sites, television and radio. The bulk of research reported through these channels is peer reviewed. A few days before a paper is published, the science journal will issue a restricted press release to qualified journalists under an agreement (called an embargo) that no one will report on the paper until a designated time. The embargo system has its flaws, but it does give reporters time to assess the research and gather expert reaction.

Contrast this with preprints. As soon as research is in the public domain, there is nothing to stop a journalist writing about it, and rushing to be the first to do so. Imagine early findings that seem to show that climate change is natural or that a common vaccine is unsafe. Preprints on subjects such as those could, if they become a story that goes viral, end up misleading millions, whether or not that was the intention of the authors.

Many reporters (including those at *Nature*) have written stories based on preprints. I'll admit that we do not yet have examples of harm from such stories, but this is probably because — at the moment — only a tiny fraction of preprints cover health-related or controversial fields.

What we have no shortage of are examples in which public understanding has been distorted by media coverage of ambiguous or just downright bad science. Take the (now retracted) paper which claimed that genetically modified maize (corn) gave rats cancer (G.-E. Séralini *et al. Food Chem. Toxicol.* **50**, 4221–4231; 2012) — its carefully orchestrated publicity campaign ensured that journalists could not seek outside scrutiny before running their news stories. The science was poor, but the claims were intriguing. Many journalists gave up the chance to vet information so as not to be late to the story.

Some scientists have pointed out that preprints are similar to conference presentations, which are also not peer reviewed. In my experience, conference proceedings often make the most misleading stories of all. Take a poster from the Pediatric Academic Societies meeting in April–May 2016, which looked at the incidence of autism and pesticide application in New York state. This was extremely early work that could not definitively establish a pattern, let alone a causal link. Still, several mainstream outlets rushed to cover it with some alarming headlines, infuriating scientists. Preprints could create similar pressures, and similarly flawed coverage — but much more often, and on a larger scale.

Another risk is the inverse — and this one could matter more to some researchers. Under the preprint system, one intrepid journalist trawling the servers can break a story; by the time other reporters have noticed, it's old news, and they can't persuade their editors to publish.

There have been cases in which a preprint that garnered news stories got a second flurry of coverage when it was published in a journal. But generally, the rule is 'it has to be new to be news'. Reacting to our blog, Tom Whipple, science editor of *The Times* in the United Kingdom, tweeted: "I'm not sure how to keep a newspaper in profitable existence that decides to give people news they've already read on the BBC."

It is not enough to shrug and blame journalists, and it is unhelpful to dismiss those journalists who can accurately convey complex science to a mass audience. Scientists need to be part of these debates — with their eyes open to how the media works. Journalists do include appropriate caveats or even decide not to run a story when conclusions are tentative, but that happens only because they have been given enough time and breathing space to assess it. If the scientific community isn't careful, preprints could take that resource away.

HOW CAN WE HAVE PREPRINTS AND SUPPORT GOOD JOURNALISM?

How can we have preprints and support good journalism? Should scientific societies or preprint advocates develop guidelines for what should and should not be posted as a preprint? Should there be a preprint moratorium on any research with public-health implications? Should universities or researchers ever publicize a preprint? Should all preprints be emblazoned with a warning aimed at journalists that work has not been peer reviewed?

Preprints could bring great prizes for science. But these questions must be brought up now, so that public understanding is not damaged as preprints flourish. ■

Tom Sheldon *is a senior press manager at the Science Media Centre (SMC) in London. Funders of the SMC include Springer Nature,* Nature*'s publisher.*
e-mail: tom@sciencemediacentre.org

© 2018 Springer Nature Limited. All rights reserved.

SEVEN DAYS *The news in brief*

PEOPLE

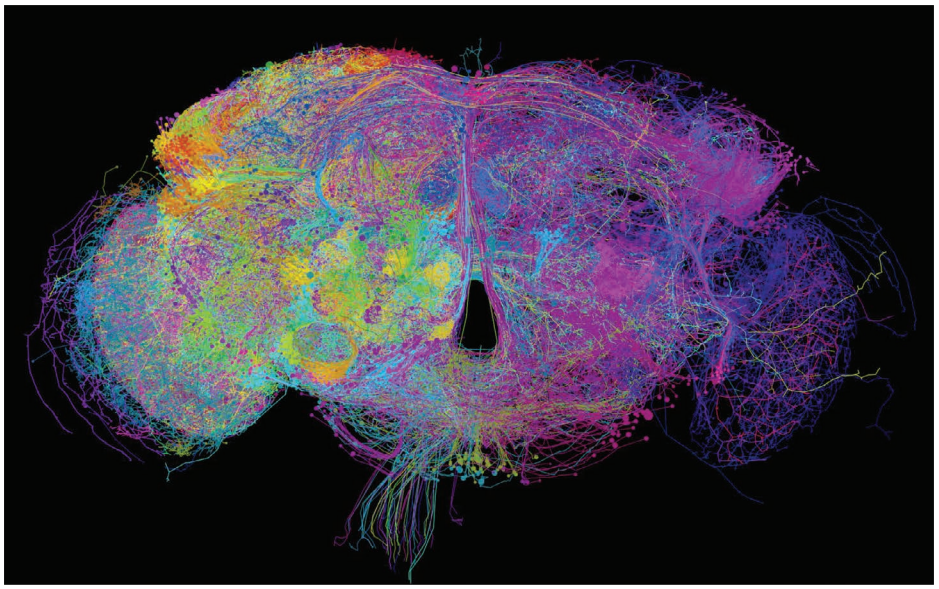
Quark finder

Physicist Burton Richter, who shared the 1976 Nobel Prize in Physics, died on 18 July, aged 87. In the 1960s, Richter helped to design a ring-shaped particle collider at the Stanford Linear Accelerator Center (SLAC) in California, which he used to discover a new subatomic particle in 1974. This particle, later named J/ψ, proved the existence of a fourth quark (known as charm) and garnered Richter the Nobel, which he shared with physicist Samuel Ting. Richter was the director of SLAC from 1984 to 1999, during which time he oversaw numerous developments in particle accelerators.

EVENTS

Three new ages

The International Union of Geological Sciences (IUGS) has declared that we are living in a geological time period called the Meghalayan age. The period covers the most recent 4,200 years of Earth's 4.6-billion-year history and is one of three ages newly named by the organization. Each is a climatically distinct period of the current Holocene epoch, which began at the end of the last ice age, about 11,700 years ago. The Meghalayan age started with a 200-year global drought recorded in sediment around the world; the official marker is a stalagmite in a cave in northeastern India. The preceding Northgrippian age was a relatively cold stage that began about 8,300 years ago. It followed the warmer Greenlandian age, which marked the start of the Holocene. Some scientists argue that the Holocene has given way to a new epoch called the Anthropocene, marked by human interference with the planet. But researchers disagree on when this proposed period would begin, and geological evidence for a new epoch is inconclusive.



VISUALISATION: PHILIPP SCHLEGEL, DROSOPHILA CONNECTOMICS GROUP, CAMBRIDGE

Fruit-fly brain mapped in epic detail

Scientists have produced a 3D image of the brain of a fruit fly (*Drosophila melanogaster*) that is so detailed, researchers can trace connections between neurons across the entire organ. The data, published on 19 July (Z. Zheng *et al. Cell* http://doi.org/gdvjwc; 2018), also revealed some neurons that have never been seen before. The scientists cut a fly's brain — roughly the size of a poppy seed — into more than 7,000 slices and shot a beam of electrons through them. A high-speed camera captured detailed pictures of each slice, generating about 21 million images that the team then combined using custom computer software. The resulting picture offers scientists a new tool with which to study the network of neurons behind fruit-fly behaviour.

Antibiotic worries

Many low-income countries lag behind in taking action to prevent bacteria from becoming resistant to antibiotics, the World Health Organization and two other United Nations agencies said in a report on 18 July. Such steps include improving sanitation and hygiene. About one-quarter of health-care facilities in Africa still lack water, the report says. And just over half of the world's population lives in countries that do not have a national system for monitoring the consumption of antibiotics. Roughly half of low-income countries do not require prescriptions for antimicrobials. Only 93 of 194 member states reported having developed strategies to stall resistance, the review finds.

Weapons opposition

More than 2,600 people and 195 organizations have signed a pledge not to support the development, manufacture, trade or use of lethal autonomous weapons. The signatories include SpaceX chief executive Elon Musk, three co-founders of Google DeepMind and thousands of researchers in the fields of computer science, robotics and artificial intelligence. In an open letter posted online on 18 July, they argue that "the decision to take a human life should never be delegated to a machine". The pledge was initiated by the Future of Life Institute, a research and outreach organization in Cambridge, Massachusetts, which focuses on the existential threats posed by technology. In August 2017, more than 100 specialists working in artificial

© 2018 Springer Nature Limited. All rights reserved.

intelligence and robotics called on the United Nations to ban autonomous weapons.

Sharing results

A platform that allows researchers to share anonymized clinical-trial data, and that already boasts results from more than 2,500 trials involving roughly 1.3 million participants, launched on 19 July. The platform, developed by the non-profit company Vivli of Cambridge, Massachusetts, will also provide tools for analysing data across different databases, and aims to capture individual patient data from academic, industry and non-profit sponsors of clinical trials. More than 2,000 of the trials currently in the database were provided by London-based pharmaceutical giant GlaxoSmithKline.



DR. SEUSS ENTERPRISES

RESEARCH

The Lorax's origin

The title character in the famed children's book *The Lorax* may have been inspired by the patas monkey (*Erythrocebus patas*), a species found across West and East Africa, researchers reported on 23 July (N. J. Dominy *et al. Nature Ecol. Evol.* **2**, 1196–1198; 2018). Theodor Seuss Geisel — better known by his pen name, Dr Seuss — wrote much of the book during a visit to Kenya in 1970. The patas monkey is native to the country, and a morphometric analysis described in the study has revealed similarities between the faces of the patas monkey and the Lorax.

Park pollution

The air in US national parks contains just as much ozone pollution as the air in many of the country's largest cities, according to a study published on 18 July (D. Keiser *et al. Sci. Adv.* **4**, eaat1613; 2018). Researchers compared ozone levels in 33 national parks and in 20 of the largest US cities between 1990 and 2014. After controlling for weather and the season, the team found that pollution levels in the parks and cities were similar. In fact, before 2000, summer ozone levels in the parks increased before they started to drop. The study authors aren't sure what caused this pattern. The findings raise health questions for park visitors because exposure to ozone pollution can exacerbate conditions such as asthma.

POLICY

Species in danger

The US government announced several proposed changes to the Endangered Species Act, a 45-year-old law that protects thousands of threatened or endangered species. The proposals, announced on 19 July, aim to alter the protocols used for determining a species' critical habitat and to streamline the process of listing and delisting organisms under the law. The US Fish and Wildlife Service, together with the National Marine Fisheries Service, would also change a current provision that automatically provides the same protections to threatened and endangered species, unless officials decide otherwise. The public has 60 days to comment on the proposals before the government finalizes them.

Science migration

A UK parliamentary group has recommended ways in which the country's immigration system for skilled workers such as scientists should change after Brexit to accommodate workers from the European Union. EU citizens currently have automatic rights to live and work in the United Kingdom, but it is not yet clear whether they will retain any of those rights after the Brexit process is complete. A report by the House of Commons Science and Technology Committee, published on 19 July and based on discussions with research organizations, recommends that the system allow skilled workers to visit visa-free for up to 180 days. Although the recommendations were designed to address the issue of how EU workers will come to the United Kingdom after Brexit, the same policies could also be applied to workers from other countries, the report says. For long-term stays, the committee suggests that the government remove an annual cap on the 'Tier 2' visa for highly skilled workers.

ENVIRONMENT

China's emissions

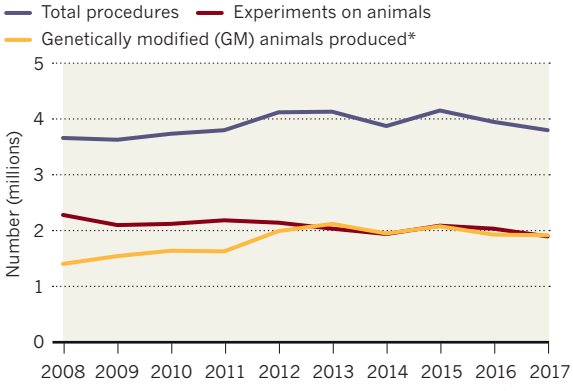
As the world's largest carbon emitter, China needs to do more if the world is to meet its climate-change goals, says a 19 July report by a former senior US energy official. David Sandalow, now at Columbia University in New York City, found that China's carbon dioxide emissions have increased by only 3% or less in the past 5 years, even as its economy has grown by 40%. The report argues that the country has enormous potential to contribute to climate-change solutions: government policies are helping to cut coal use and increase use of low-carbon alternatives, and the nation is investing in technologies such as electric vehicles and solar power. But some of its policies — such as converting coal to synthetic natural gas and financing coal power plants around the world — run counter to these goals, the report concludes.

TREND WATCH

The production of genetically modified animals has levelled off in UK laboratories, after a decade that saw a 37% rise in the number bred. Last year, 1.9 million of these animals were produced, about the same as the year before, according to annual statistics released by the UK government on 19 July. In addition, 1.9 million experimental procedures, including those used for basic research, drug studies and surgical training, were conducted on animals last year — a 7% drop since 2016 and a 17% decline since 2008.

RESEARCH ANIMALS

The number of animals used in UK labs in 2017 fell by 4% — mostly due to the 7% decline in experimental procedures on animals.



SOURCE: UK HOME OFFICE

NATURE.COM
For daily news updates see:
www.nature.com/news

© 2018 Springer Nature Limited. All rights reserved.

NEWS IN FOCUS

PHILIPPINES Government lures scientists home to bolster research workforce **p.453**

PUBLISHING German scientists cut off from journals in row over open access **p.454**

COSMOLOGY Planck telescope's final map marks end of an era **p.455**



MALARIA The race to defuse a ticking time bomb in southeast Asia **p.458**

JONATHAN WEBB



Northern-quoll populations have declined by more than 75% since the introduction of poisonous cane toads, which the quolls mistake for food.

ECOLOGY

Evolution experiment aims to save Australian marsupial

If it succeeds in quolls, the conservation strategy could also help protect Tasmanian devils.

BY APRIL REESE

On an island off Australia's north-central coast, researchers are conducting an unprecedented experiment: mixing endangered animals that have evolved genetic defences against their biggest foe with those that haven't, in the hope that their offspring will take after the wiser parent.

The subject of the experiment is one of Australia's most imperilled marsupials, the northern quoll (*Dasyurus hallucatus*). This squirrel-sized carnivore is struggling to survive a decades-long onslaught of poisonous and invasive cane toads, which quolls mistake for prey, with devastating results. The team has tested the match-making technique in captive-bred quolls, and reported the results last month (E. Kelly and B. L. Phillips *Conserv. Biol.* http://doi.org/csf2; 2018).

The researchers are now testing the technique in wild animals on Indian Island and, if the experiment is successful, it could provide some of the first real-world evidence that targeted gene flow — which involves pushing an adaptive trait through an at-risk population to boost resiliency — could be used to save an endangered species.

Australia, which has the world's highest mammal extinction rate, is desperate for new conservation tools, says Euan Ritchie, a wildlife ecologist at Deakin University in Melbourne. The quoll research is exciting, he says, because it "provides hope we can still achieve conservation wins, even for the most challenging and pressing of situations". The real test will be the ongoing wild trial. If targeted gene flow works, the ▶

© 2018 Springer Nature Limited. All rights reserved.

research team says it could help other threatened species, too — including Tasmanian devils and corals on the Great Barrier Reef.

In the 80 years since agriculture officials introduced the cane toad (*Rhinella marina*) to northeastern Australia to control a sugarcane-devouring beetle, the amphibians have spread across the state of Queensland, the Northern Territory and large chunks of Western Australia. Their rapid advance has devastated northern-quoll populations, which have shrunk by more than 75%.

Ecologists Ella Kelly and Ben Phillips knew from their previous research that some quoll populations in Queensland had developed an aversion to the toads over the years (E. Kelly and B. L. Phillips *Behav. Ecol.* **28**, 854–858; 2017). The researchers, both at the University of Melbourne, wondered whether the trait could be successfully bred into vulnerable quoll populations that cane toads hadn't yet reached. That could make those 'naive' quolls more resilient to toad invasions, if enough animals in a given group have the trait.

To test this idea, the scientists bred animals in captivity, mixing northern quolls from a toad-infested area of Queensland that displayed an aversion to toads with naive quolls from a toad-free island in the Northern Territory. Kelly and Phillips then exposed the resulting offspring to a toad leg to gauge whether the young quolls recognized the threat. They found that most of the young quolls wouldn't touch the toad legs.

The finding suggests that the trait is inherited, rather than taught by mother quolls, and may be dominant, the researchers say. "That's the first hurdle that needs to be jumped in showing targeted gene flow," Phillips says. "Without a genetic basis, there is no point in introducing [toad-smart quolls] into the population. And we found there is a genetic basis."

The captive-quoll study is an important step towards demonstrating that targeted gene flow is a viable strategy to aid quoll conservation, says Sarah Fitzpatrick, a conservation biologist at Michigan State University's Kellogg Biological Station in Hickory Corners. "Many behaviours are plastic, and therefore are not necessarily controlled by certain genes," she says. "If this were the case for toad-eating behaviour, targeted gene flow would not work."

Buoyed by the results of the captive study, Kelly and Phillips, along with University of Melbourne colleague Chris Jolly, decided in May 2017 to see whether toad-smart and naive quolls would produce toad-smart offspring in the wild. They released 54 quolls on toad-infested Indian Island — a mix of naive Northern Territory quolls, toad-averse quolls from Queensland and hybrid offspring.

When the researchers returned in April this year to check on the quolls, they found good and bad news. Many fewer quolls survived than the team had anticipated — just 16 animals, according to the researchers' population estimate. But encouragingly, the group included offspring that seemed to be toad-smart, which would suggest that they had inherited the trait from their parents. The team is now analysing genetic samples taken from the survivors.

Kelly and Phillips plan to return to the site again next April, to see how the remaining quolls fare.

As the experiment continues, the pair is seeking permission from wildlife officials in Western Australia to introduce northern quolls with the toad-smart trait into populations in the path of the cane-toad diaspora to gird them against future invasions. "It would be a tragedy not to try," Phillips says. ■

ASTRONOMY

NASA aims for Sun's corona

The Parker Solar Probe will make humanity's closest approach to its home star.

BY ALEXANDRA WITZE

Step aside, Icarus: NASA has made a spacecraft that can fly through the Sun's atmosphere without melting.

On 6 August, if all goes to plan, the US$1.5-billion Parker Solar Probe will lift off from a launch pad at Florida's Cape Canaveral. Just three months later, it will whizz much closer to the Sun than any spacecraft has ever come, to take the first-ever direct measurements of the star's maelstrom of energy.

But that's just the beginning. Over the next 7 years, the craft will loop around the Sun another 23 times, passing nearer and nearer — ultimately flying about 6.2 million kilometres above the surface, well within the solar corona. That's nearly seven times closer than the record mark set by the German Helios 2 spacecraft in 1976.

The Parker Solar Probe aims to answer some of the biggest outstanding questions about the Sun, such as how its corona is heated to millions of degrees while the surface beneath it stays relatively cool[1]. The spacecraft will also visit the birthplace of the solar wind, a flood of energetic particles that streams out into the Solar System at speeds of up to 800 kilometres a second. When the solar wind slams into Earth, it generates beautiful polar aurorae, but it can also disrupt satellite communications and navigation systems.

"We're going to be right where all the interesting stuff happens," says Nicola Fox, a solar physicist at the Johns Hopkins University Applied Physics Laboratory (APL) in Laurel, Maryland, and the mission's project scientist.

Data from the deep-diving probe should allow researchers to improve their understanding of the complex picture of how particles, magnetic fields and energy combine in the Sun. "This is going to be such a game-changer," says Nicholeen Viall, a solar physicist at NASA's Goddard Space Flight Center in Greenbelt, Maryland.

MORE ONLINE

TOP NEWS



Humans are altering seasonal climate cycles worldwide go.nature.com/2a3rm3v

MORE NEWS

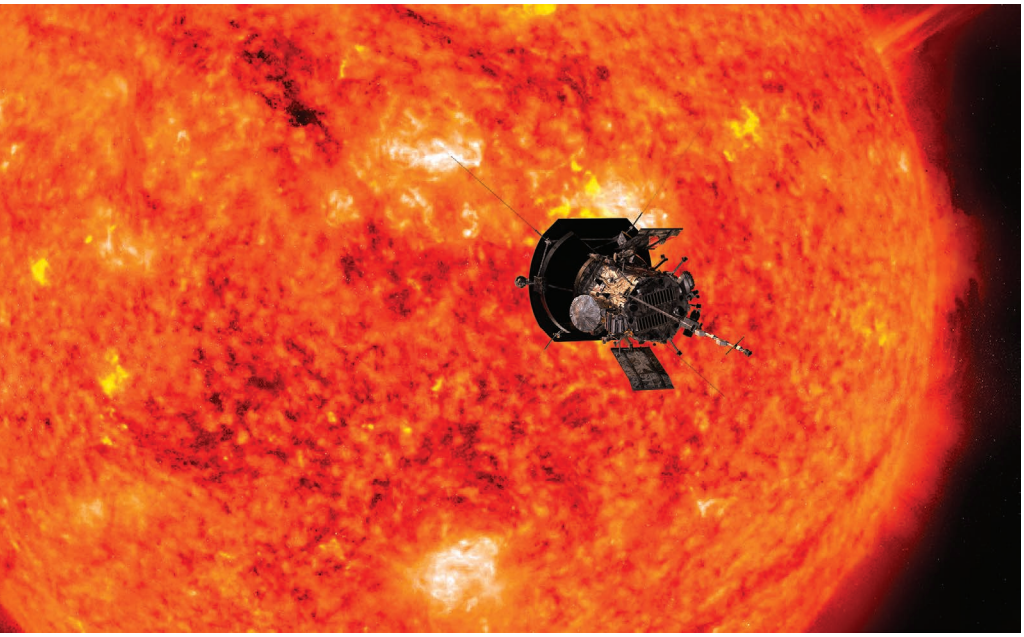
- 3D image reveals hidden neurons in fruit-fly brain go.nature.com/2lhma4s
- Why computer science needs its own gatekeeper go.nature.com/2lvwllv
- The monkey that could be Dr Seuss's Lorax go.nature.com/2mfgyeu

NATURE PODCAST

Automata through the ages; problems with pet DNA tests; and a conservation conundrum nature.com/nature/podcast

DAVE SANDERS/*THE NEW YORK TIMES*/REDUX/EYEVINE

© 2018 Springer Nature Limited. All rights reserved.



The Parker Solar Probe will travel seven times closer to the Sun's surface than any previous spacecraft.

JHUAPL

Space physicists have dreamt of a mission that would fly through the solar corona, or at least travel inside the orbit of Mercury, the innermost planet, since 1958. In the same year, Eugene Parker — the physicist at the University of Chicago in Illinois for whom the probe is named — first proposed the existence of the solar wind[2].

After decades on the drawing board, the mission is finally approaching launch. Eight weeks after lift-off, it will fly past Venus, using the planet's gravity to slow down and slip into a tighter orbit around the Sun. Five weeks after that, on 3 November, the probe will make its first close approach — at more than 24 million kilometres, or 35 times the solar radius, from its surface.

From there, the spacecraft will loop around the Sun, drawing gradually closer as it flies past Venus six more times. That trajectory will give the probe ample time to gather data, says Yanping Guo, an engineer at APL who designed the mission's trajectory.

Somewhere between the first close approach (at 35 solar radii) and its final ones (within 10 solar radii) the probe will encounter the Alfvén surface, a boundary at which the solar wind becomes supersonic. Inside the Alfvén surface, the Sun's magnetic field dominates; outside, the solar wind is more detached and streams away on its own.

Crossing that boundary with a spacecraft will be similar, symbolically, to the moment when the Voyager 1 probe entered interstellar space in 2012, says Justin Kasper, a physicist at the University of Michigan in Ann Arbor who has studied the Alfvén transition[3]. The moment will mark humanity's passage to another realm in the Solar System. "I'm confident that something special will happen," he says.

The Parker Solar Probe bristles with an array of instruments designed to sample the corona directly. Protecting them is a 2.4-metre-wide heat shield made of 11-centimetre-thick carbon foam sandwiched between layers of carbon composite. It can withstand temperatures of nearly 1,400 °C. The solar panels that power the spacecraft will be kept cool with a water-tubing system similar to a car's radiator. During the searing conditions of close approach, most of the solar panels will fold back to shelter in the heat shield's shade.

Mission scientists hope that the Parker Solar Probe will kick off a new era of studying the Sun. In 2020, the European Space Agency plans to launch its Solar Orbiter spacecraft, which will study the Sun at higher latitudes and from a more distant point in space than the Parker Solar Probe will. Also by 2020, the Daniel K. Inouye Solar Telescope will come online in Hawaii, where it will make daily maps of the solar corona.

For his part, the 91-year-old Parker is looking forward to seeing the waves and turbulence in the solar wind — which he predicted — measured by the probe that bears his name. "I expect to find some surprises," he says. ■

1. Fox, N. J. *et al. Space Sci. Rev.* **204**, 7–48 (2016).
2. Parker, E. N. *Astrophys. J.* **128**, 664–676 (1958).
3. Kasper, J. C. *et al. Astrophys. J.* **849**, 126 (2017).

FUNDING

Philippines sweetens deal for scientists who return home

But some academics say more needs to be done to train and retain early-career researchers.

BY ANDREW SILVER

A renewed government effort to draw Filipino scientists back to the Philippines by paying them to set up their own research labs or teach has met with a mixed reception.

Some Filipino researchers applaud the goals of the effort. The government says it needs to bring research expertise back home to solve some of the country's most pressing problems, such as climate-change mitigation.

Others suggest that resources would be better directed to mentoring early-career scientists before they think about leaving. "If you want Filipino scientists, you grow them," says Vena Pearl Bongolan, an applied mathematician at the University of the Philippines Diliman in Quezon City.

An existing project designed to address the issue is the Balik Scientist Program, established in the 1970s. Since 1993, it has offered research grants for up to three years, as well as round-trip airfare and duty-free equipment imports, to Filipino scientists willing to return home.

Last month, President Rodrigo Duterte signed a law that instructs the Department of Science and Technology to allocate more money to the Balik programme by adding benefits for returning scientists, such as a monthly housing allowance, medical insurance and assistance for researchers' children to attend schools of their choice. Between 2007 and April 2018, 207 scientists joined the Balik programme, some for stints of a couple of months and others for several years. ▶

© 2018 Springer Nature Limited. All rights reserved.



JULIAN ABRAM WAINWRIGHT/GETTY

The Philippine government wants to boost the number of agricultural scientists in the country.

The programme cost 173 million Philippine pesos (US$3.2 million) from 2007 to the end of 2017.

The department has not yet established exactly how the changes to the programme will work and how much extra money it will get. Several sectors are listed as high-priority recruitment areas, including space, energy, artificial intelligence and agriculture and food.

GROWING DIASPORA

In 2013, the Philippines had only 187.7 scientists per million people, one of the lowest densities of researchers in the region. The latest figures from the Department of Science and Technology estimate that the number of Filipino science and technology workers who moved overseas jumped from 9,877 in 1998 to 26,917 in 2013.

The revamped Balik programme is part of a government plan to increase development in the country, which came into effect last year. The government aims to boost the scientific workforce to 300 researchers per million people by 2022.

Biologist Michael Velarde, a current Balik scientist at the University of the Philippines Diliman, says that the extra support for the programme is a good idea because the country needs more overseas-trained researchers to address challenges including preventing the spread of diseases such as Zika and studying how climate change affects health.

And he thinks the added financial incentives will attract more talent back to the Philippines.

But Bongolan is frustrated that the country is paying to relocate scientists from abroad rather than offering more scholarships that encourage local students to pursue science at university.

She participated in the Balik programme in July 2008 after studying and working in the United States for a decade, and is sceptical about whether some participants are in the Philippines for long enough to contribute meaningful research and development.

The Department of Science and Technology did not respond to *Nature*'s request for details on what participants had achieved. But a department spokesperson said each scientist sets goals for their time in the country, and that all participants so far have met their targets. A statement posted on the department's website last month noted that the Balik programme had "significantly contributed to the acceleration of the scientific, agro-industrial and economic development of the country".

Miguel Garcia, a Filipino neuroscience and economics PhD student at the University of Zurich in Switzerland, says that the Philippines needs more than money to lure researchers back from overseas. Scientists need access to the right facilities and other researchers in their discipline; these are lacking in his field, he says.

Garcia also thinks the government should be encouraging scientists of any nationality to work in the Philippines. "Why tie your research interests based on someone's nationality when the government could set up the kind of research it needs, and attract scientists — regardless of nationality — to do it?" he asks. ■

PUBLISHING

Mega-publisher cuts off German scientists

Negotiations with Elsevier have stalled over open access.

BY HOLLY ELSE

Elsevier last week stopped thousands of scientists in Germany from reading its recent journal articles, as a row escalated over the cost of a nationwide open-access agreement. The move comes just two weeks after researchers in Sweden lost access to the most recent Elsevier research papers, when negotiations on a contract there broke down over the same issue.

Negotiators in Germany and Elsevier now seem to be waiting for the other to blink, says Joseph Esposito, a publishing consultant in New York City. The highly public nature of the stand-off means that "any deal Elsevier does with them becomes the de facto deal for the entire world", he adds.

Elsevier's move to cut off some German researchers also provides a test as to whether the scientists can survive without a subscription deal with the mega-publisher, says Ralf Schimmer, director of scientific information at the Max Planck Digital Library in Munich, Germany. "If it comes to hardship and misery, then the negotiators might be forced back to the negotiating table." His organization provides journal access to the dozens of Max Planck Institutes and their libraries, and its contract with Elsevier finishes at the end of this year.

Elsevier declined to comment on the move, which was reported by negotiators, some affected libraries and Germany's national university association. The Amsterdam-based company instead reiterated a 5 July statement saying it was committed to reaching a deal with the German consortium Projekt Deal, which is brokering an agreement on behalf of hundreds of Germany's universities and research organizations. Projekt Deal declared on 6 July that it had suspended talks with Elsevier. The publisher produces more than 2,500 journals, which issue in excess of 400,000 papers each year.

© 2018 Springer Nature Limited. All rights reserved.

SUBSCRIPTION SHUTDOWN

The stand-off was sparked by talks between Elsevier and Projekt Deal, which is pushing to create a collective subscription agreement to replace the individual deals each institution has held with the publisher.

This new type of deal would offset the cost of publishing under open-access terms against the price paid for subscriptions to paywalled journals. These 'read and publish' contracts have become popular in recent years, as governments in some European countries have tried to make the fruits of publicly funded science open to all.

Academic library consortia in Austria, the Netherlands, the United Kingdom, Sweden and Finland have all struck read-and-publish deals with various publishers — including Wiley, Springer Nature and Taylor and Francis — that cover varying parts of their portfolios. (*Nature*'s news team is editorially independent of its publisher, Springer Nature.) Last month, the Massachusetts Institute of Technology in Cambridge became the first US institution to enter into one, covering journals published by the UK Royal Society of Chemistry. But some national consortia are now coming up against fierce resistance to such contracts from Elsevier.

In May, talks collapsed between Elsevier and the Swedish Bibsam Consortium, which brokers deals on behalf of 85 institutions across the country. Their existing contract expired on 30 June, and some researchers in Sweden have now lost access to all Elsevier journal articles published after this date.

TOUGH COMPROMISE

Negotiators in Germany and Sweden want all their papers published in Elsevier journals to be open access as part of any new contracts. They have said that they will not pay more than they did previously for subscriptions. But, until now, the Dutch publisher has offered other countries read-and-publish deals that cover only a small proportion of a country's publishing output.

Apart from a brief shutdown in early 2017, which affected about 70 German institutions, Elsevier has provided mostly uninterrupted access to German institutions whose contracts have expired, while negotiations continue. Around 200 are thought to be affected by the latest switch-off, according to Projekt Deal.

The affected universities and research institutes can still source missing Elsevier articles through inter-library loans from the 150 or so institutes whose contracts have not yet expired.

The pressure on Elsevier to accept a read-and-publish contract is increasing, says Bernhard Mittermaier, head of the central library at the Jülich Research Centre in Germany and a member of the Projekt Deal negotiating team. He sees the widespread shift to 'gold' open access — whereby journals make papers freely available once published — as inevitable. "Publishers would be well advised to take it into their own hands and flip by themselves," he says. "Science in Germany will not break down."

Esposito thinks that the German institutes have leverage over Elsevier because their researchers can access papers on the illicit sharing site Sci-Hub — but that does not necessarily mean that Elsevier will back down.

"The Germans insisted on conducting this negotiation in public," he says, which could lead to a no-deal scenario because other countries might now want the same. If the stand-off continues without resolution, Elsevier will be closely monitoring journal article submissions from researchers based in Germany, Esposito predicts. "If it drops sharply, Elsevier will likely reconsider its position." ■

EARLY UNIVERSE

Big Bang telescope finale is end of an era in cosmology

Collaboration behind Europe's Planck mission releases its final maps of the early Universe.

BY DAVIDE CASTELVECCHI

A transformative era in cosmological science ended last week when the European Space Agency's Planck telescope released its final maps of the early Universe. Planck was the last in a line of three major space telescopes to study the cosmic microwave background (CMB), the faint afterglow of the Big Bang, resulting in the most precise measurements yet of the age, geometry and composition of the cosmos. With space agencies in Europe and the United States hesitant to fund a follow-up mission, Planck looks set to be the last CMB-focused satellite for many years — marking a big change for cosmologists.

"There's a whole generation of young scientists who grew up with Planck," says cosmologist Jan Tauber, the mission's project scientist with the European Space Agency (ESA) in Noordwijk, the Netherlands.

For more than two decades, scores of ground-based and balloon-borne experiments have also studied the CMB. They have largely focused on mapping minute variations in the CMB's temperature across the sky to create charts of the Universe that have become the gold standard of cosmology. Planck, which collected data from 2009 to 2013, helped researchers to pin down the age of the Universe (about 13.8 billion years), its geometry (essentially flat) and its composition (95% dark matter and dark energy). In particular, the latest release solidifies an earlier prediction based on Planck data that the Universe should be expanding 9% slower than is currently observed (see go.nature.com/2jt0sbi).

"There's a whole generation of young scientists who grew up with Planck."

The temperature maps and the science they produced have been a "great achievement" but they don't have much more to give, says Peter Coles, a theoretical cosmologist at Maynooth University in Ireland who is not part of the collaboration.

PLANCK DIASPORA

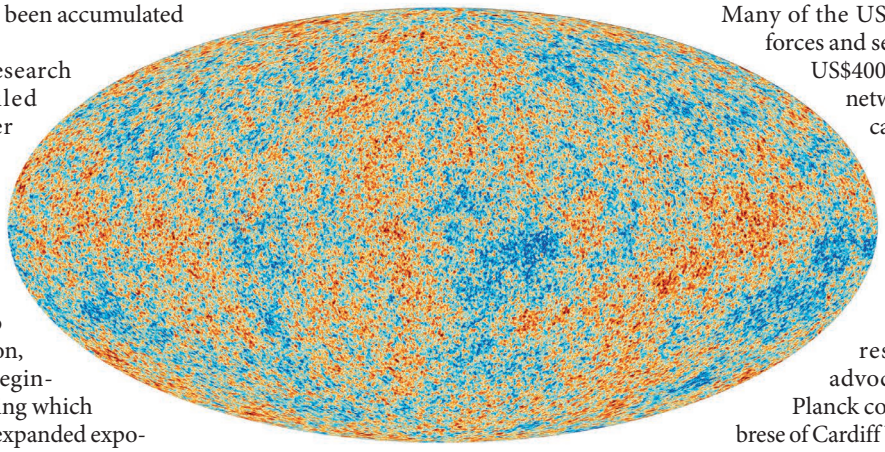
Many scientists who worked on the mission have already moved on to other projects. Silvia Galli, who help to lead the latest study, became part of Planck in 2013 after her PhD and is one of the few dozen scientists left. Now, she says she will probably join many of her colleagues who are working on Euclid, a major European mission to map the Universe's galaxies on an unprecedented scale that is preparing for launch in 2021. Euclid is an optical telescope, not a microwave detector, which makes it a technically different kind of mission. On a personal level, it is exciting to move on to new endeavours, she says.

But the prospect of no major CMB mission in the pipeline worries many researchers. "Scientifically, it would be a disaster," says Galli, who is based at the Institute of Astrophysics in Paris (IAP). "There is a risk that a lot of know-how ▶

© 2018 Springer Nature Limited. All rights reserved.

▶ and expertise that had been accumulated will be lost."

The focus of CMB research is now to make detailed measurements of other parameters, including polarization: a slight tendency for the microwaves' electromagnetic fields to align in particular directions. In this, researchers hope to find a signature of inflation, the brief period at the beginning of the Big Bang during which the Universe would have expanded exponentially. Planck mapped polarization — as did a NASA telescope in the 2000s — but with limited sensitivity. "Only 10% of the information in the polarization has been exploited," says Jacques Delabrouille, an astrophysicist at the University of Paris Diderot who helped to design Planck. "The CMB still has plenty of secrets to yield."



ESA/PLANCK COLLABORATION

The Universe's cosmic microwave background measured by the Planck satellite.

THE POLARIZATION PROBLEM

NASA and ESA have so far declined to fund big new satellites to study the CMB — although several US groups are working on ground and balloon instruments to measure polarization. One reason for space agencies' reticence to fund big CMB projects is that experiments have yet to find a polarization signature of inflation. An experiment called BICEP2 claimed to have detected that signature in 2014, but Planck data later showed that it was just dust in the Milky Way. In the latest studies, the Planck collaboration also looked for an inflation signature but didn't detect one. However, "the fact that we haven't seen it yet doesn't mean that it's not there", says cosmologist Richard Gott of Princeton University in New Jersey.

Many of the US teams are now joining forces and seeking funding to build a US$400-million, next-generation network of ground telescopes called CMB-S4, which will be much more sensitive to the inflation signature than anything that has come before. That's the next big thing on the horizon, Tauber says.

And many CMB researchers continue to advocate for space missions. Planck cosmologist Erminia Calabrese of Cardiff University, UK, is pushing for Europe to join LiteBIRD, a proposed Japanese probe that would look for the signature of cosmic inflation while keeping costs down. From space, it would have the advantage of seeing the entire sky — something that cannot be done from the ground.

Other cosmologists are trying to jump-start CMB-polarization research in Europe. Davide Maino, a senior Planck member at the University of Milan in Italy, is collaborating on an Italian-led polarization experiment that will be based on the Spanish island of Tenerife. If ground-based experiments see even a hint of inflation, that will encourage space agencies to fund major missions, researchers say. ■

© 2018 Springer Nature Limited. All rights reserved.

THE ENEMY IN WAITING

Malaria is a ticking time bomb in southeast Asia. Scientists are now racing to stamp it out before unstoppable strains spread.

BY AMY MAXMEN



In Siem Pang, Cambodia, a young man with malaria is treated.

© 2018 Springer Nature Limited. All rights reserved.



© 2018 Springer Nature Limited. All rights reserved.



Malaria has been on the rise in northeastern Cambodia, where people live in open homes in rural, mosquito-ridden areas.

PHOTOGRAPHS: ADAM DEAN FOR *NATURE*

Throngs of men and women ventured into the forests of northeastern Cambodia in April, lured by a bumper crop of a rare tropical treat called samrong. After days of hiking through the wilderness, some of the travellers returned to their homes with a bounty of the wrinkled seeds, which fetch a high price as a special dessert or an ingredient in traditional Chinese medicine. But many soon fell ill. Khong Chhoem, a 56-year-old rice farmer, says the fevers hit him a few days after the expedition. His muscles hurt. His eyes hurt. He had unbearable nightmares. A health worker told Chhoem that he had tested positive for *Plasmodium falciparum*, the deadliest kind of malaria-causing parasite. But because a wave of malaria was sweeping through the region, medicine was in short supply. Chhoem eventually found a shop that carried the drugs he needed, and he recovered. But in the intervening days, mosquitoes probably sucked up the parasites in his blood and spread them to other people.

After years in decline, malaria infection rates seem to be on the rise in northeastern Cambodia, where people are moving deeper into lush, mosquito-ridden territories in search of timber and seasonal goods such as samrong (*Scaphium affine*). Their movements provide opportunities for *P. falciparum* — which requires both human and insect hosts — to thrive. There are other contributors as well, such as treatment delays that allow the parasites to linger and spread, and an alarming decline in the potency of gold-standard malaria drugs called artemisinin-based combination therapies (ACTs).

What happens next here matters for the entire world; malaria remains one of the biggest killers in low-income countries. Estimates of the number of deaths each year range from 450,000 to 720,000 — and ACT pills keep that toll from being much higher. And although southeast Asia accounts for just 7% of malaria cases worldwide, it has a notorious history as the breeding ground for strains of malaria parasites that resist every drug thrown at them and then spread to other regions.

In 2015, reports of drug resistance prompted the governments of five countries in the Greater Mekong Subregion — Cambodia, Thailand, Vietnam, Laos and Myanmar — to pledge to banish *P. falciparum* from the region by 2025. Together with the World Health Organization (WHO), the countries drew up plans and budgets. This year, the nations' governments have committed US$41 million towards the effort; the Global Fund to Fight AIDS, Tuberculosis and Malaria also backed elimination efforts in the region, with a 3-year, $243-million grant. Donors such as the Bill & Melinda Gates Foundation and the Asian Development Bank will add more than $20 million to the fight this year.

But the rise of cases in northeastern Cambodia shows how difficult getting to zero will be — and how crucial. As long as *P. falciparum* exists, it can resurge. And the last parasites remaining are the hardest to find. They reside in the hinterlands, borderlands and war zones

© 2018 Springer Nature Limited. All rights reserved.



Near Cambodia's border with Laos, people often travel on small tractors to harvest timber and other goods from the forest, where they can contract malaria.

— and in people who show no signs of the disease. "Malaria is very clever — it hides out where you don't know and comes back when you aren't ready," says Ladda Kajeechiwa, a malaria-programme manager at a branch of the Mahidol Oxford Tropical Medicine Research Unit (MORU) in Mae Sot, Thailand.

To eliminate malaria, public-health officials are trying to cover the region with volunteers and health workers who can dispense malaria drugs promptly, and report any signs of an upsurge. They are also looking to researchers for tools that can forestall drug resistance, technologies that improve detection of the parasite, and evidence on strategies for rooting it out. The pressure is on. "Everybody has fears of malaria resurgence," says Myaing Nyunt, a malaria researcher at Duke Global Health Institute in Durham, North Carolina. She worries that donors will tire of the cause, and that the people in the Greater Mekong region will grow to resent the obsession with malaria when they have more pressing needs. Yet if the effort lets up, malaria will roar back.

"That is why it is so important to eliminate it as quickly as possible," Myaing Nyunt says.

THE RISE OF RESISTANCE

Southeast Asia has served as a launch pad for drug-resistant malaria multiple times, with disastrous consequences. In the 1950s, Cambodian authorities reported that chloroquine, the main cure for malaria at the time, had stopped clearing some infections. Within three decades, chloroquine-resistant strains had spread to Africa[1], where 90% of the world's malaria cases occur, and the death toll started to climb. By 2000, malaria was killing about one million people a year globally. Doctors said it was like watching a couple of Boeing 747 planes crash every day.

"Malaria is very clever — it hides out where you don't know and comes back when you aren't ready."

Around 2005, the trend started to reverse as ACTs became available around the world. The main ingredient in these treatments, artemisinin, rapidly wipes out the majority of parasites, and a longer-acting partner drug in the combo mops up the rest. When ACTs are working as they should, the number of parasites detectable in a finger-prick of blood drops by half within three hours of taking the medicine.

But in 2008, researchers in western Cambodia reported[2] that parasite numbers in some patients were not falling as quickly as they ought to. To monitor the situation, scientists throughout the region

© 2018 Springer Nature Limited. All rights reserved.



A technician in Siem Pang, Cambodia, prepares blood cells for shipment to Bangkok, where biologists will study the malaria parasites living in the cells.

started shipping blood samples to high-tech labs in cities including Bangkok, Singapore and London. It was through such work that researchers uncovered a secret weapon that parasites had gained against artemisinin[3] — mutations in a gene called *Kelch 13*. They found that several *Kelch 13* mutations had popped up in the Greater Mekong Subregion, and the share of parasites harbouring such mutations was rising. Resistance to artemisinin didn't cause a rise in deaths, partly because of the partner drugs. But some of those drugs had begun to falter, too. One lineage of *P. falciparum* from Cambodia was found to contain both a powerful *Kelch 13* mutation and a gene duplication that helps it to survive treatment by the partner drug piperaquine[4]. Last October, molecular biologists reported[5] that this parasite had spread to Thailand, Laos and Vietnam. Now, some members of that family also resist another partner drug, mefloquine[6] (see 'A crucible for resistance').

Drug resistance has a way of growing from a nuisance to a serious problem. Although an incompletely cleared infection might not kill a person, the parasites that survive can go on to infect others and their lineage grows stronger. To stay a step ahead of resistance, researchers at MORU renovated a dilapidated clinic in Siem Pang in northeastern Cambodia earlier this year. They are expanding a trial to test a combination of artemisinin with two partner drugs, rather than one. That combination should kill parasites that haven't yet made it to the triple-resistant stage. But trial leader Arjen Dondorp, deputy director of MORU in Bangkok, says that the only certain way to halt resistance

"All the investments we've made so far here and in Africa will amount to an incredible waste of money."

is to rid the Mekong region of malaria altogether. At the moment, the WHO considers elimination feasible because there are relatively few infections in southeast Asia. But the opportunity to do so won't last long: as ACTs start failing, case numbers will rise. "This is the scenario we predicted and now we see it unfolding in Cambodia," Dondorp says.

To hasten the elimination of *P. falciparum* in the region, public-health officials in the five countries have promised to strengthen the basic mechanisms of malaria control, which include maintaining a steady force of health workers. But gaps remain in this system, particularly in Cambodia and Myanmar, which are among the poorest countries in Asia. To help them, Benjamin Rolfe, head of the Asia Pacific Leaders Malaria Alliance in Singapore, has been garnering funds and political support throughout Asia. "We are

© 2018 Springer Nature Limited. All rights reserved.

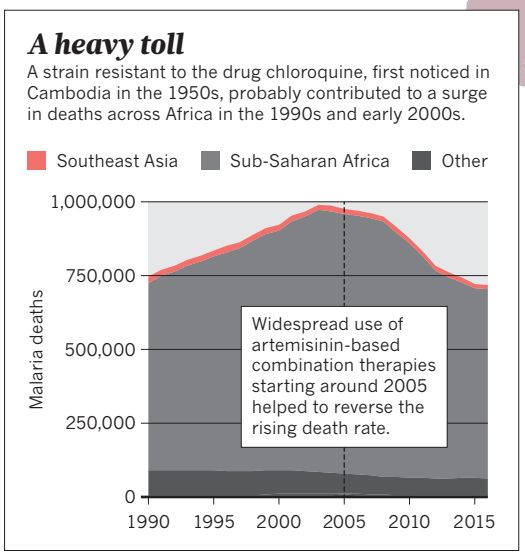


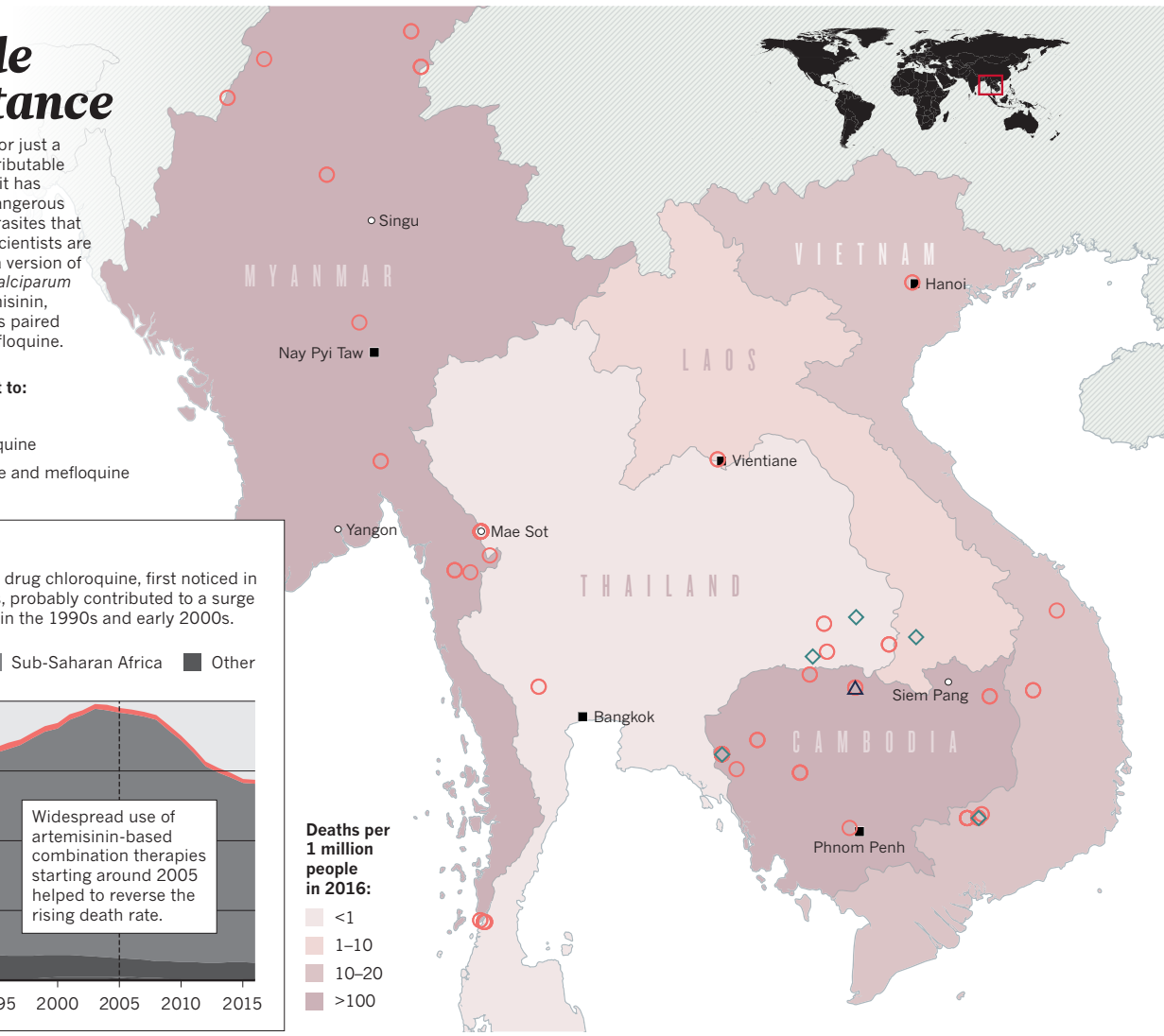
A crucible for resistance

Southeast Asia accounts for just a few per cent of deaths attributable to malaria worldwide, but it has repeatedly given rise to dangerous drug-resistant malaria parasites that spread to other regions. Scientists are paying close attention to a version of the parasite *Plasmodium falciparum* that resists the cure artemisinin, along with some drugs it is paired with: piperaquine and mefloquine.

Malaria parasites resistant to:

- ○ Artemisinin
- ◇ Artemisinin and piperaquine
- △ Artemisinin, piperaquine and mefloquine





SOURCE: INSTITUTE FOR HEALTH METRICS AND EVALUATION; WORLDWIDE ANTIMALARIAL RESISTANCE NETWORK; REFS 4–6

so close to the end game that it is not good business sense to let drug resistance happen," Rolfe says. "If we do, all the investments we've made so far here and in Africa will amount to an incredible waste of money."

A SCOURGE IN HIDING

Recent findings suggest that the task of elimination will be even harder than expected. Health workers typically diagnose malaria using a simple rapid diagnostic test (RDT): a pin-prick of blood is added to a test strip, which quickly turns red if it detects proteins from *P. falciparum*. The tests adequately distinguish malarial fevers from other infections. But Mallika Imwong, a molecular biologist at Mahidol University in Bangkok, wondered whether some people might carry too few parasites to make them ill — or to show up on a RDT. To find out, she fished for parasite DNA in vials of blood shipped from Cambodia, Myanmar and Vietnam, using a sensitive method called real-time polymerase chain reaction (PCR). In 2015, she reported[7] that only one-quarter of *Plasmodium* infections were detected by the usual RDT on average. The disparity was at times remarkable. "I was really surprised," says Imwong. "In one village, only 4% or 5% of the people were positive with RDTs, but 68% were infected."

Researchers have not yet discovered how or why the parasites can persist at low levels in a person's bloodstream. Whatever the reason, there are signs[8] that these covert microbes multiply when conditions are right — such as when the rainy season starts.

Knowing that silent infections exist is one thing, but actually finding them is quite another. Health officials say it is unrealistic to conduct PCR analyses on vials of blood collected across the Greater Mekong Subregion. Rather, health workers need something like a rapid diagnostic test that requires no refrigeration, lab equipment or special training, but that can detect parasites even when they are not abundant.

In central Myanmar, a roving team of doctors, researchers and technicians are piloting one such technology. Thura Htay, a young physician and public-health researcher swerves on his motorcycle over rugged red-dirt roads with a green plastic supply bag gripped between his thighs. He is pensive, overseeing his group on its mission to collect pin-pricks of blood from 357 people in 3 remote villages by the end of the day. In these mountains, the things most likely to thwart the mission are flash floods and signs of wild elephants. (When startled by the sound of motorbikes, the animals sometimes attack people, homes and farms.)

Today, the path remains clear and the team arrives at the Shwe Kyin village before noon. A few dozen people linger on a concrete platform, where tables, chairs and bowls of pickled tea leaves, fried nuts and sweets have been neatly arranged in anticipation of the team's visit. Although no one from the village has been diagnosed with malaria this year, Thura Htay says, "I want to know if asymptomatic malaria is here".

A teacher in a rhinestone-studded top holds out her finger to a team member. He pricks it and places a drop of blood onto a conventional diagnostic test, and then onto an experimental version that the team is assessing to see whether it is more sensitive. Then he pricks another

© 2018 Springer Nature Limited. All rights reserved.



Forests along the Mekong river and its tributaries provide a breeding ground for mosquitoes carrying malaria.

finger and presses it against two kinds of filter paper — one intended to capture DNA, and the other, proteins. The team swiftly hangs the papers on a clothes line to dry.

The teacher, Khin Zar Lwin, says she's happy to help scientists learn about her village, but the emphasis on malaria by this and other groups feels like overkill; few people here get the disease, and there are other pressing health concerns. Pregnant women and elderly people who need urgent care, for example, must travel for hours on foot and then by boat down a tumultuous river to reach the nearest hospital, in the town of Singu.

In the late afternoon, the team packages the tests and blood spots into coolers strapped to the back of motorcycles, and returns to a concrete building in Singu. Team members hang the papers in a pair of cupboards, cooled and screened off to protect against mould and flies. Thura Htay sleeps on a bed beside them to ensure they are safe.

Every few weeks, the team ships samples to a renovated lab at the government's Department of Medical Research in Yangon. There, researchers compare the results of the two rapid diagnostic tests with results from a PCR analysis on blood spots collected on one of the filter papers.

By the end of this year, Thura Htay — and his boss, Myaing Nyunt — will know whether the new diagnostic tests can be used to map out the regions where symptomatic and asymptomatic malaria occur. Such 'hotspot' maps will help officials to decide on elimination strategies for different regions. But the question remains of how to handle communities with hidden parasites. One possible answer comes from eastern Myanmar.

"We work day by day, year by year, without guarantees on what happens tomorrow."

A few years ago, researchers mapped the malaria hotspots in Karen, a state along Myanmar's border with Thailand that has resisted government rule since the country gained independence from Britain seven decades ago. (A symbol of that is its rejection of the state's official name, Kayin.) Conflict has made many residents mistrustful of outsiders. But when MORU's Ladda Kajeechiwa visited one of the hotspots in May this year, an old man in the village of Ta Oo Hta greeted her warmly.

The man, Bo Lay, recalls the toll that malaria took during the community's darkest days. Burmese troops and armed groups attacked Ta Oo Hta several times, he says, and burnt it to ashes at least once. Each bout of violence drove people into the mountains, where they slept among the snakes and leeches. Those who became

© 2018 Springer Nature Limited. All rights reserved.



severely ill with malaria trekked through the jungle and crossed the Moei River to refugee clinics along the Thai border in Mae Sot. "Sometimes they died along the way," he says.

Bo Lay was glad when he heard about MORU's plan to wipe out malaria, and he discussed it with his neighbours, easing their suspicions of outsiders asking to draw blood and deliver drugs. This plan consisted of controlling malaria — and then pressing the accelerator on elimination. Beginning in 2014, MORU and other local organizations strengthened a network of 1,222 village malaria workers by training and paying them to treat and report any malaria cases that they find. In addition, they analysed blood samples across the state with PCR to find pockets of malaria. They then asked every healthy person in a hotspot — about 4% of Karen's population — to swallow ACTs once a month for three months[9]. In surveys after the intervention, the researchers found that the incidence of malaria in the state had dropped by 90%. The only stronghold of the disease is in a rugged region in the north, where armed separatist groups rarely accept outside interventions.

"We have strong evidence that elimination can happen," says François Nosten, a French researcher who founded MORU's facility in Mae Sot, the Shoklo Malaria Research Unit, in 1986. But to his frustration, governments in the region have put mass-drug-administration strategies on hold for a number of reasons.

For instance, it takes a great deal of work to convince people who don't feel ill to take medicines that can cause fatigue and nausea, and it requires funds to pay and organize health workers to survey populations for the parasite and distribute drugs. If a significant number of people don't take the full course of therapy, parasite reservoirs remain — and some researchers fear that this situation could foster drug resistance.

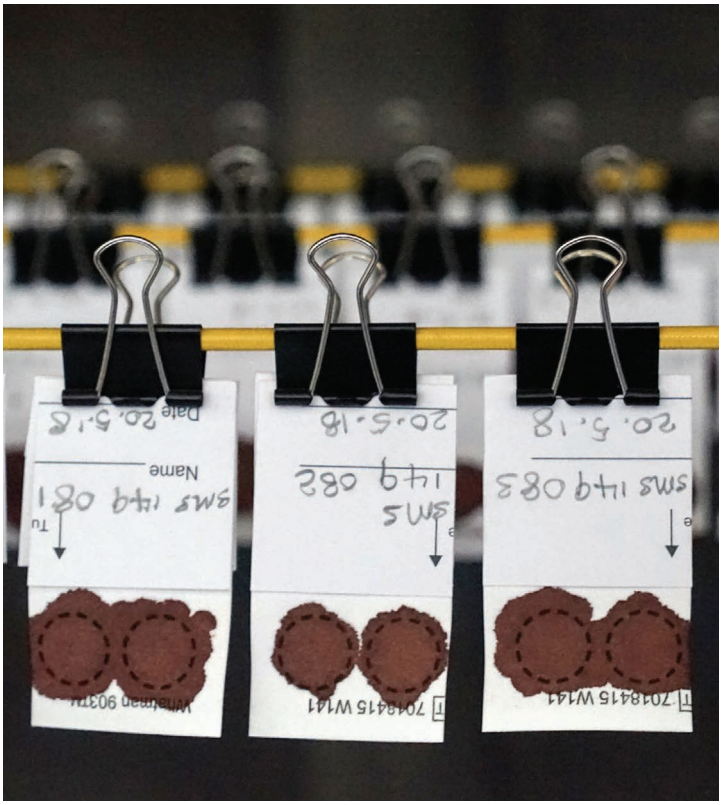
Karen's nearly disease-free status remains tenuous as long as malaria exists in the broader region. And its defence against imported cases relies on continued international aid because there is no stable health system in Karen. At the moment, the Wellcome Trust, the Global Fund and the Bill & Melinda Gates Foundation provide funds for malaria workers that ensure that any cases that pop up in the state are treated before the parasites have a chance to spread. But if donors put their funds elsewhere — or fighting breaks out again in Karen, as it has in the past — the system will fall apart and malaria will resurge. "The NGOs [non-governmental organizations] could pull out, the political situation is unstable," Kajeechiwa says. "We work day by day, year by year, without guarantees on what happens tomorrow."

A similar fragility exists in other states along Myanmar's border, such as Kachin in the north and Rakhine in the west, the site of the ongoing Rohingya crisis. Since August 2017, more than half a million members of the Muslim Rohingya ethnic group have fled killings, arson and rape at the hands of Myanmar's troops and nationalistic Buddhists. In these areas of the country, aid organizations and private clinics handle malaria because the public-health system is fractured. That concern resonates with Saw Nay Htoo, programme director of the Burma Medical Association, a non-profit organization working with minority ethnic groups. In an office in Mae Sot where vines penetrate the panelled ceiling, he says, "we want to eliminate malaria, but first we need peace".

THE LAST MILE

Myaing Nyunt believes the common enemy of malaria could help to unite warring factions. She manifests that hope. A political dissident who fled from Myanmar's military government in 1988, Myaing Nyunt now works with government researchers there. Squeezed into the back seat of a jeep beside Zaw Lin, an official in Myanmar's Vector-Borne Disease Control department in Nay Pyi Taw, she explains why she decided to partner with researchers who are part of a government that has been accused of human-rights violations. Because there is no quick fix for malaria — no powerful vaccine — the fight against the disease requires a sustainable health system. As a scientist, she says she can develop top-notch malaria-elimination approaches and tools, but those won't matter if governments do not put them into practice.

Always frank, Nyunt says that the decision to work closely with



Blood spots collected in Myanmar will be used to validate diagnostic tests that detect the malaria parasite with more sensitivity than current technology.

AMY MAXMEN

the government has made her research slow and hard. There are bureaucratic meetings to attend, labs to be built from scratch and students who are in need of long-term mentorship and stipends. "We are talking about a health system that has been underfunded and underdeveloped since the 1960s, and that's not going to change overnight," she says. "But you have to do it this way because changes must come from within."

If the new diagnostic that she and Thura Htay are road-testing in central Myanmar works, Zaw Lin will help the government to decide whether to pre-emptively distribute ACTs in hotspots, as was done in Karen, or to routinely screen people in those places with the new test, and treat those who test positive. He doesn't shy away from talking about the country's conflicts. In addition to political instability, he and other public-health officials find themselves pushing to sustain enthusiasm for their mission amid other health concerns. But he has seen at first hand how malaria bounces back if it's not obliterated. It resurged in Myanmar in the 1990s, just as it seems to be doing in Cambodia and several other countries now.

Losing inertia in the Mekong region, the motherland of drug resistance, could be a disaster. "This is the last-mile challenge in front of us," Zaw Lin says. "If we don't eliminate malaria, deaths can shoot up like a hidden bomb." ■

Amy Maxmen *writes for* Nature *from San Francisco, California.*

1. von Seidlein, L. & Dondorp, A. *Expert Rev. Anti Infect. Ther.* **13**, 715–730 (2015).
2. Noedl, H., Se, Y., Socheat, D. & Fukuda, M. M. *N. Engl. J. Med.* **359**, 2619–2620 (2008).
3. Ariey, F. *et al. Nature* **505**, 50–55 (2014).
4. Imwong, M. *et al. Lancet Infect. Dis.* **17**, 491–497 (2017).
5. Imwong, M., Hien, T. T., Thuy-Nhien, N. T., Dondorp, A. M. & White, N. J. *Lancet Infect. Dis.* **17**, 1022–1023 (2017).
6. Rossi, G., De Smet, M., Khim, N., Kindermans, J.-M. & Menard, D. *Lancet Infect. Dis.* **17**, 1233 (2017).
7. Imwong, M. *et al. Malar. J.* **14**, 381 (2015).
8. Nguyen, T.-N. *et al. Lancet Infect. Dis.* **18**, 565–572 (2018).
9. Landier, J. *et al. Lancet* **391**, 1916–1926 (2018).

© 2018 Springer Nature Limited. All rights reserved.

COMMENT

POLICY Pet genomics is oversold and under-regulated **p.470**

ROBOTS Responses to automata over millennia reveal familiar fears **p.473**

VIROLOGY From literature search to vaccine candidate without a lab **p.477**

POLITICS Government shutdowns of the Internet have a health impact **p.477**



SIMON DAWSON/BLOOMBERG/GETTY

An engineer inserts a lithium-ion battery from an electric vehicle into a testing system for recycling.

Ten years left to redesign lithium-ion batteries

Reserves of rare metals used in electric-vehicle cells are dwindling, so boost research on iron and silicon alternatives, urge **Kostiantyn Turcheniuk** and colleagues.

Electric vehicles need powerful, light and affordable batteries. The best bet is commercial lithium-ion cells — they are relatively compact and stable. But they are still too bulky and expensive for widespread use.

The performance of rechargeable lithium-ion batteries has improved steadily for two decades. The amount of energy stored in a litre-sized pack has more than tripled, from around 200 watt hours per litre ($Wh\,l^{-1}$) to more than 700 $Wh\,l^{-1}$. Costs have fallen by 30 times, to around US$150 per kilowatt hour (kWh). But that still exceeds the $100 per kWh goal for affordability set by the US Department of Energy. And batteries that are powerful enough for an electric car (50–100 kWh) still weigh around 600 kilograms and take up 500 litres of space.

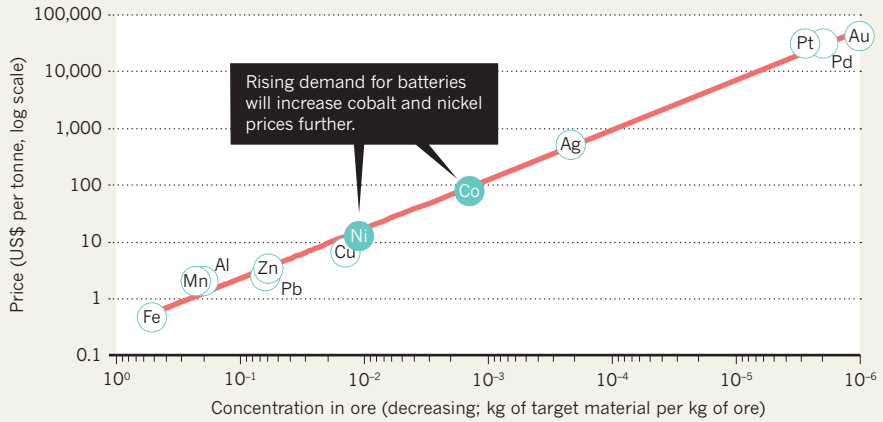
The pace of advance is slowing as conventional technology approaches fundamental limits. The amount of charge that can be stored in gaps within the crystalline structures of electrode materials is nearing the theoretical maximum. Projected market growth will not lower prices substantially — the markets are already large.

Worse, the materials used in electrodes, notably rare metals such as cobalt and nickel, are scarce and expensive. Surging battery production has almost quadrupled wholesale prices of cobalt over the past two years, from $22 to $81 per kilogram. ▶

© 2018 Springer Nature Limited. All rights reserved.

METAL PRICES

The high cost of extracting metals from low-concentration ores pushes up their market price.



SOURCE: LONDON METAL EXCHANGE

High demand and prices are already encouraging some producers to cut corners and violate environmental and safety regulations. For example, in China, dust released from graphite mines has damaged crops and polluted villages and drinking water[1]. In Africa, some mine owners exploit child workers and skimp on protective equipment such as respirators. Small artisanal mines, where ores are extracted by hand, often flout laws. Some companies, including BMW, follow strict policies to verify their cobalt suppliers[2]. Many do not.

Alternative types of electrode based on cheap, common metals such as iron and copper need to be developed urgently. In our view, the most promising candidates involve 'conversion materials', such as copper or iron fluorides and silicon. These store lithium ions by bonding chemically with them. But the technology is still at an early stage. Problems with stability, charging speed and manufacture must be overcome.

We call on materials scientists, engineers and funding agencies to prioritize the research and development of electrodes based on abundant elements. Otherwise, the roll-out of electric cars will stall within a decade.

SCARCE AND EXPENSIVE

Lithium-ion batteries work by shuffling lithium ions between two electrodes. Ions flowing from the anode to the cathode discharge a current, which powers the car. The lithium ions flow back when the battery is recharged.

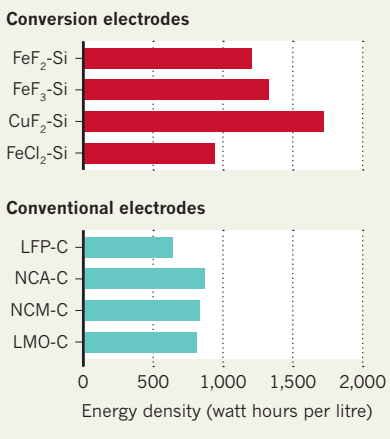
In commercial cells used today for electric vehicles, the lithium ions are held in tiny voids within the crystals that make up the electrodes (these are known as intercalation electrodes). The anodes are typically made from graphite and the cathodes from metal oxides. Common oxides include lithium nickel cobalt aluminium oxide (NCA, commonly $LiNi_{0.8}Co_{0.15}Al_{0.05}O_2$) or lithium nickel cobalt manganese oxide (NCM, often $LiNi_{0.6}Co_{0.2}Mn_{0.2}O_2$ or $LiNi_{0.8}Co_{0.1}Mn_{0.1}O_2$). A lithium-ion car battery with a 100 kg cathode requires 6–12 kg of cobalt and 36–48 kg of nickel.

The prices of metals reflect demand, supply and the costs of extracting them from ores. Cobalt is pricey because it is rare and highly sought after. It requires capital-intensive processes to produce it, involving roasting, flash smelting and the consumption of poisonous gases[3]. Cobalt is often a by-product of copper and nickel mining, and can also need separating from other metals.

Few cobalt mineral deposits are concentrated enough to be worth mining. Most deposits contain just 0.003% of the metal; more than 0.1% is needed to achieve prices of $100 to $150 per kg (ref. 4). Production costs jump for poor ores because more rock must be processed (see 'Metal prices'). Thus, only 10^7 tonnes of cobalt out of 10^{15} tonnes potentially available in Earth's crust are profitable to extract[5]. Similarly, only 10^8 of 10^{15} tonnes of nickel reserves are commercially viable[5].

ENERGY ADVANTAGE

Batteries that use conversion electrodes can store more energy in a given unit stack volume than those using conventional electrodes.



C, carbon (graphite); LFP, lithium iron phosphate; NCA, lithium nickel cobalt aluminium oxide; NCM, lithium nickel cobalt manganese oxide; LMO, lithium manganese oxide.

Cobalt-rich minerals are found in just a few places[6]. The Democratic Republic of the Congo (DRC) supplied more than half (56%) of the 148,000 tonnes of the metal mined worldwide in 2015 (ref. 6). Most of this goes to China, which holds stockpiles of 200,000 to 400,000 tonnes[6]. Australia hosts 14% of the world's cobalt reserves but has yet to exploit them fully. Cobalt has been extracted from the deep sea floor, but mining here would be too expensive, ecologically and economically.

Likewise, nickel production is dominated by a dozen nations. In 2017, Indonesia, the Philippines, Canada, New Caledonia, Russia and Australia together supplied 72% of the 2.1 million tonnes mined globally. Of this, less than one-tenth went to batteries; the rest was used mainly in steel and electronics. Nickel is cheaper to extract than cobalt, through a series of reactions with hydrogen and carbon monoxide[7]. Nonetheless, rising demand has boosted nickel prices by about 50% since 2015, from $9 to $14 per kg.

Both cobalt and nickel have suffered sudden price hikes and crashes. For example, disrupted Australian supplies, increased demand from China for steel and speculation by hedge-fund managers led to a five-fold surge in the price of nickel and a tripling of that for cobalt in 2008–09.

PROJECTED SHORTFALLS

If nothing changes, demand will outstrip production within 20 years. We expect this to occur for cobalt by 2030 and for nickel by 2037 or sooner.

Car manufacturers and governments project that 10 million to 20 million electric cars will be produced each year by 2025. If each car battery requires 10 kg of cobalt, by 2025, electric vehicles would need 100,000–200,000 tonnes of cobalt per year — most of the world's current production. Similarly, 400,000–800,000 tonnes of nickel would be required annually, or 20–40% of all the metal used today. More would be needed when trucks, buses, aeroplanes and ships become battery-powered.

SOURCE: REF. 9

By 2050, producing 50 million to 80 million electric vehicles a year would require 500,000–800,000 tonnes of cobalt. Beyond 2030, this would far exceed current mining capacities. Similarly, 2–3 times more nickel would be needed by 2050. Nickel shortages would be evident by the mid-2030s.

Recycling cannot replenish supplies. Lithium-ion batteries last for 15–20 years, 3 times longer than the 5–7 years for lead-acid batteries. Refiners might exploit poorer quality ores, especially as prices climb. But greater processing costs would push the prices higher. Once supplies peak, we estimate that the price of an electric car battery could leap by more than $1,000. 'Peak cobalt' might be delayed by a few years if cathodes can be made using less metal. But

© 2018 Springer Nature Limited. All rights reserved.



MICHAEL ROBINSON CHAVEZ/THE WASHINGTON POST/GETTY

A miner enters a copper and cobalt mine in Kawama, Democratic Republic of the Congo.

cathode materials that use less cobalt, which are under development, degrade quicker and would need to be replaced more often.

ABUNDANT MATERIALS

The answer is to make lithium-ion battery electrodes from common metals, such as iron and copper. A kilogram of iron, for example, costs just 6–9 cents. The global iron reserves stand at 76 billion tonnes: easily enough to supply the billions of batteries that could replace today's petroleum car engines.

Already, Chinese electric buses use conventional cathodes made from iron-rich materials such as lithium iron phosphates (LFPs; $LiFePO_4$), which can survive many cycles of charging. But they are less efficient at holding charge than cobalt- and nickel-based cathodes (see 'Energy advantage'). Twice as many iron-based cells[8,9], at nearly twice the cost, are needed to drive the same distance.

Equally inefficient are commercial electrodes made from another common metal, manganese (such as lithium manganese oxide or LMO, commonly as $LiMnO_2$ or $LiMn_2O_4$). Earlier this year, a cathode (based on lithium manganese niobium oxyfluoride) was demonstrated in the lab that could almost match the energy storage capacity of cobalt- and nickel-based materials[10]. But a high voltage is needed to charge it, making it unsafe for use in vehicles.

The most promising alternative, in our view, is to use conversion materials in electrodes. Copper and iron fluorides and silicon react with the lithium ions. A transition metal in a conversion cathode can host up to six times more lithium atoms than one in a standard cathode. Such materials swell more to accommodate the extra lithium atoms. Cell designs must allow for this distortion.

Future anodes are likely to be based on silicon. It is produced from sand and stores nearly ten times more lithium ions by mass than graphite does. Combining conversion cathodes with silicon anodes in the next generation of lithium-ion battery cells could allow cells to store more than twice as much energy as the best conventional ones by volume, and more than three times by weight[8,9]. Half as many cells would be required to power electric vehicles, also halving costs, weight and volume.

Silicon-anode technology is advancing. Tesla already uses small amounts of silicon in anodes made of graphite for lithium-ion cells in its electric vehicles, and BMW announced plans to incorporate silicon-dominant anodes in its future electric vehicle batteries. Other companies, too, are developing silicon-rich anode materials. These include the Californian firms Enevate in Irvine, Enovix in Fremont and Sila Nanotechnologies in Alameda (G.Y. is a board member, shareholder and chief technology officer of Sila Nanotechnologies).

Metal fluoride cathode technology has yet to move beyond the lab[9]. Conversion-type batteries can take 20 hours to power up; this needs to be reduced to tens of minutes. They also require up to one-third more energy to charge; this should be no more than 10%. And their stability needs to be improved, from 5–500 cycles to 1,000–2,000 cycles.

Undesirable interactions need to be minimized between conversion materials and electrolyte substances. And the microstructures and compositions of electrodes and electrolytes must be optimized. New architectures need to be explored that contain the swelling and shrinkage of electrodes while maintaining their conductivity.

NEXT STEPS

Several technological breakthroughs are needed to secure the future of affordable battery-powered transport.

First, cell performance needs to improve. Materials scientists need to produce high-performance fluoride materials for the ▶

© 2018 Springer Nature Limited. All rights reserved.

electrodes. Electrochemists should develop more-effective electrolytes. Engineers need to develop tools for manufacturing these materials.

Government agencies and leading car manufacturers should fund this research, which will require billion-dollar investments. In our view, the best way to apportion this money is by targeting projects addressing key battery challenges, as the US Department of Energy's Advanced Research Projects Agency-Energy (ARPA-E) does. Developing batteries free from cobalt and nickel should be one priority.

As promising battery materials and cell technologies emerge, funding should be refocused towards improving their characteristics and viability. Production processes and costs must be addressed. We expect that synthesizing conversion electrode materials will require different steps, including forming certain nanoscale structures. Treatments involving series of solutions and gases, for example, might be borrowed from other sectors such as food, pharmaceuticals, filtration and composite manufacturing.

Lithium-ion battery manufacturers have already invested billions of dollars in dozens of 'giga-factories' to boost the electric-vehicle market. With co-ordination, thought and planning, these can be set on a new path to deliver the next generation of affordable batteries. ■

Kostiantyn Turcheniuk *is a research scientist at the School of Materials Science & Engineering;* **Dmitry Bondarev** *is an undergraduate student at the School of Chemical and Biomolecular Engineering;* **Vinod Singhal** *is a professor of business at the Scheller College of Business; and* **Gleb Yushin** *is a professor of materials science at the School of Materials Science & Engineering, Georgia Institute of Technology, Atlanta, USA.*
e-mail: gleb.yushin@mse.gatech.edu

1. Whoriskey, P. In Your Phone, In Their Air. *The Washington Post* (2 October 2016).
2. Amnesty International. *Time to Recharge: Corporate Action and Inaction to Tackle Abuses in the Cobalt Supply Chain* (Amnesty International, 2017).
3. Freeman, G. K .W. Process for recovery of cobalt by selective precipitation of cobalt-calcium double salt. US patent US6264904B1 (2001).
4. British Geological Survey. *Cobalt Commodity Profile* (BGS, 2009).
5. US Geological Survey. *Mineral Commodity Summaries 2018* (USGS, 2018).
6. Brown, T. J. *et al. World Mineral Production 2011–15* (British Geological Survey, 2017).
7. *Nature* **59**, 63–64 (1898).
8. Nitta, N., Wu, F., Lee, J. T. & Yushin, G. *Mater. Today* **18**, 252–264 (2015).
9. Wu, F. & Yushin, G. *Energy Environ. Sci.* **10**, 435–459 (2017).
10. Lee, J. *et al. Nature* **556**, 185–190 (2018).

G.Y. declares competing financial interests; see go.nature.com/2ntpjme.

Pet genomics medicine runs wild

Genetic testing for dogs is big business. It is too easy for companies to sell false hope, warn **Lisa Moses**, **Steve Niemi** and **Elinor Karlsson**. They call for regulation.

Last year, a 13-year-old dog, let's call her Petunia, started having trouble walking and controlling her bladder and bowels. Distressed, her owners bought a US$65 genetic test through a direct-to-consumer (DTC) company. It suggested that the pug carried a mutation that is linked to a neurodegenerative condition similar to the human disease amyotrophic lateral sclerosis (ALS, or motor neurone disease).

What published data there are[1,2] suggest that as few as 1 in 100 dogs that test positive for this common mutation will develop the very rare disease, and Petunia's condition was also consistent with more-treatable spinal disorders. But her owners chose to put her to sleep, convinced that she would otherwise suffer progressive and irreversible paralysis and eventual death.

Genetic testing for pets is expanding. Hundreds of thousands of dogs have now been genetically screened, as Petunia was, and companies are beginning to offer tests for cats. But the science is lagging. Most of these tests are based on small, underpowered studies. Neither their accuracy nor their ability to predict health outcomes has been validated. Most vets don't know enough about the limitations of the studies, or about genetics in general, to be able to advise worried owners.

Pet genetics must be reined in. If not, some companies will continue to profit by selling potentially misleading and often inaccurate information; pets and their owners will suffer needlessly; and opportunities to improve pet health and even to leverage studies in dogs and cats to benefit human health might be lost. Ultimately, people will become more distrustful of science and medicine.

WEAK SCIENCE

Global spending on pets has grown by 14% over the past 5 years (see 'Animal lovers'), and worldwide annual spending on pet care is currently estimated to be about $109 billion.

Genetics is one of the newest additions to this booming industry. Worldwide, at least 19 laboratories are now marketing genetic-testing products. Some vets use the results to help to diagnose sick pets, or to recommend that healthy ones be subject to expensive and sometimes invasive tests, such as bone-marrow biopsies. Some dog breeders use the tests to try to reduce the incidence of inherited diseases. At least one US veterinary hospital chain is now recommending genetic testing for all dogs, saying that the results allow "individualized healthcare" and can guide behavioural training.

We believe that three major problems plague pet genetic testing in its current state.

Lack of validation. In both humans and animals, mapping genetic variants to risk of disease is incredibly challenging. But most dog genetic tests are based on studies of candidate genes, which is a problematic approach. In such studies, researchers test a handful of human or animal genes for mutations that both match an expected inheritance pattern and seem likely to be pathogenic, for instance because they affect the structure of an expressed protein.

In humans, fewer than 2% of candidate-gene studies have stood up[3] to further investigations using more-advanced methods, such as genome-wide association studies. As a result of these shortcomings, geneticists who study humans must now bring more evidence to designate a genetic variant as 'disease-causing'. Through collaborative efforts involving industry, academia, physicians and patients, every clinical variant identified is now scored on a five-point scale, from 'pathogenic' to 'benign'[4,5]. (Clinical variants are those that are linked to medically important phenotypes.)

No such careful reassessment has occurred in veterinary medicine. Many of the 200 tests offered by companies are based on only a single small candidate-gene study (see, for example, go.nature.com/2nquntx).

© 2018 Springer Nature Limited. All rights reserved.



FRANCK ROBICHON/EPA/REX/SHUTTERSTOCK

Spending on pet care is increasing around the world.

The data obtained are rarely made public. Neither is the information required to contextualize the results. Providers who have followed up with owners to check the accuracy of their tests' health predictions have reported case studies involving a small number of dogs[6], but nothing on the scale needed to provide statistically reliable results[7].

Imprecise results or interpretation. A recent analysis of DTC results supplied by human genetic-testing companies found that, of all genetic variants presented in people's test results, 40% were inaccurate[8]. Other variants were reported by DTC companies to increase disease risk, but are considered benign by independent labs. Finally, the companies failed to screen for all known disease-linked mutations for any given gene, giving an incomplete picture of genetic susceptibility to disease.

The same problems apply to the DTC products now offered for pets. Dogs, for example, are frequently screened for mutations in the gene *ABCB1* that are associated with dangerous sensitivity to several common drugs. Three different mutations have been associated with this phenotype, but the documentation provided by genetic-testing companies implies that they test for only one. Thus, a dog declared 'clear' for a given gene might still harbour other known, clinically relevant mutations in that gene that the company has not tested for.

Conflicts of interest. In the absence of effective industry self-policing or government regulations, there are many potential conflicts of interest associated with making profits from pet owners. As an example, a pet health-care corporate database could be used to identify the breed of an owner's dog. The corporation owning the database might then notify the clinic (which might be owned by the same company) that a genetic test for a specific illness is warranted, regardless of that test's medical value. If the test comes back positive, the clinic's vet might recommend preventive steps, such as specific pet foods (made by the same company), periodic screening tests (performed by the company's clinical lab), and more-frequent exams (performed at the company's vet clinics), even though there may be low or no risk of disease in the first place.

TSUNAMI OF DATA

The companies do provide multiple caveats about interpreting genetic test results in their terms and conditions. Yet their warnings are not as prominent as their claims. And companies tend to be vague about which variant they test for, even in publications[6].

All of this is worrying because veterinary medicine is about to be hit by a tsunami of genomic data. Currently, pet owners can purchase the tests directly from companies and obtain results that ostensibly report on risk for more than 100 different diseases for less than $200, and some kits provide a 'health report' to take to the vet. Within the next five years, owners will have the option of whole-genome sequencing for pets of any species, using the same technology used for human medicine. Such results could accelerate genetic-based diagnoses and even the genetic editing of pets.

In principle — with large, well-powered studies, involving perhaps tens of thousands of animals — genetic testing could be used to predict the risk of common diseases affecting millions of pets, including cancer, heart disease, diabetes and epilepsy. The use of pets in preclinical and clinical research to benefit humans is growing. For certain diseases, such as cancer and diabetes, researchers are sometimes using pet dogs as models in preference to lab rodents. Although it is harder to control environmental variables and to obtain high numbers of subjects, the use of dogs offers the advantages of spontaneous disease models that might be closer to the human version. Additionally, pet owners voluntarily enrol their pets and continue to feed and care for them at home. Thus, many of the costs and regulations that govern the

© 2018 Springer Nature Limited. All rights reserved.



A cheek swab is all that's needed for a genetic test.

ELISE AMENDOLA/AP/REX/SHUTTERSTOCK

same species when they are used in the lab can be bypassed. Such methodological shifts make it imperative that the science of pet genetics is robust.

Largely because current regulatory mechanisms for pets focus on treatments, not diagnostics, there is no oversight of genetic testing for pets in the United States or in the European Union. To our knowledge, neither the US Food and Drug Administration (FDA) nor the European Medicines Agency (EMA) has proposed quality standards, or sought public input about the tests' pros and cons.

To bring the untamed wilderness of pet genetic testing under control, we propose five steps.

Establish standards. All stakeholders, from pet owners and the companies providing the tests to researchers and regulators, should work together to establish standards both for testing methodology and for the reporting of test results. Samples should be collected, stored, shipped and analysed in specified ways; the number of animals should exceed a certain threshold to support claims made about a specific test; and reports should meet specified criteria. Reports must be sufficiently clear, with information about potential caveats provided at a predetermined point in the document and in easy-to-read language. Organizations such as the World Small Animal Veterinary Association could take the lead on convening meetings with this goal.

Create guidelines. A select group of stakeholders — which would comprise working groups — could develop guidelines that specify the standards that should be adopted. Such a document could be similar to the ethics guidelines[9] for the use of novel technologies in lab-animal research developed by the British Small Animal Veterinary Association science subcommittee. Ideally, the instructions would have global reach, and as pet genetic testing grows, such guidelines might eventually need to become law.

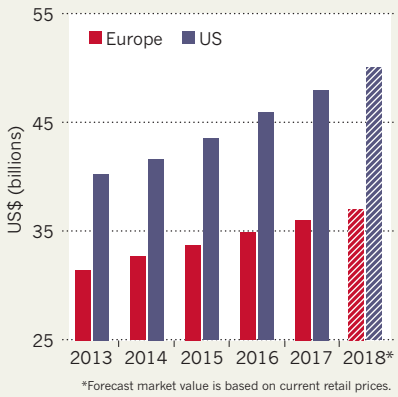
Share data. Existing pet genetic databases, which are produced by industry, academia and government agencies, should be shared among all parties, with steps taken to ensure that individual pets and owners can't be identified. Companies might be incentivized to share data if the proposed guidelines require sample sizes that can be achieved only through pooling data. To improve transparency, stakeholders can draw on the knowledge accumulated across the past two decades by scientists researching human genetics, who have been grappling with issues around data sharing and patient privacy.

Recruit tools and expertise. Bioinformaticians, computer scientists trained in machine learning, and others with expertise in big data will be needed to manage and analyse the tremendous volume of incoming information. Such specialists might be persuaded to enter the burgeoning field of pet genetics because of the potential financial gains. Or they could be lured by the possibility of pet genetics offering an inroad into solutions to human disease.

Educate counsellors. A cadre of pet genetic counsellors needs to be established from among vets and others working in animal care. These professionals would provide support and advice to pet owners following genetic tests and could potentially be affiliated with their counterparts in human genetic counselling at leading academic medical centres.

ANIMAL LOVERS

In the United States and Europe, spending on pet care has steadily grown over the past six years.



*Forecast market value is based on current retail prices.

SOURCE: EUROMONITOR

DATA ETHICS

Other issues might require ongoing public dialogue and the appointment of expert advisory boards. For instance, should for-profit pet-testing firms have unfettered access to all the results they generate, for internal corporate use or to sell to other companies engaged in pet care — as is the case now? Can an owner, who might want to breed their pet and sell the offspring, keep any results secret (as they can currently) or should they be required to divulge this information to buyers?

In the United States alone, some 70% of households own pets. Done right, the use of genetic testing in companion animals could be a powerful way to better connect people to the possibilities of genetics for treating disease. Done wrong, it could erode trust in science for an increasingly sceptical public. ■

Lisa Moses *is a research scholar at the Center for Bioethics, Harvard Medical School, Boston, Massachusetts, USA; and a practising veterinarian in the Pain and Palliative Care Service at the MSPCA-Angell Animal Medical Center, Boston, Massachusetts.* **Steve Niemi** *is a veterinarian and director of the Office of Animal Resources, Harvard University, Cambridge, Massachusetts.* **Elinor Karlsson** *is assistant professor at the University of Massachusetts Medical School, Worcester, Massachusetts; and director of vertebrate genomics at the Broad Institute of Harvard and MIT, Cambridge, Massachusetts.*
e-mails: lisa_moses@hms.harvard.edu; sniemi@fas.harvard.edu; elinor.karlsson@umassmed.edu

1. Coates, J. R. & Wininger, F. A. *Vet. Clin. North Am. Small Anim. Pract.* **40**, 929–950 (2010).
2. Zeng, R. *et al. J. Vet. Intern. Med.* **28**, 515–521 (2014).
3. Ioannidis, J. P. A., Tarone, R. & McLaughlin, J. K. *Epidemiology* **22**, 450–456 (2011).
4. Richards, S. *et al. Genet. Med.* **17**, 405–424 (2015).
5. Harrison, S. M. *et al. Genet. Med.* **19**, 1096–1104 (2017).
6. Donner, J. *et al. PLoS Genet.* **14**, e1007361 (2018).
7. Goldstein, B. A., Navar, A. M., Pencina, M. J. & Ioannidis, J. P. A. *J. Am. Med. Inform. Assoc.* **24**, 198–208 (2017).
8. Tandy-Connor, S. *et al. Genet. Med.* https://doi.org/10.1038/gim.2018.38 (2018).
9. Yeates, J. W. *J. Small Anim. Pract.* **57**, 67–73 (2016).

E.K. declares competing interests; see go.nature.com/2mwyv2b.

© 2018 Springer Nature Limited. All rights reserved.



PHILIPPE HURLIN/MUSEE DES ARTS ET METIERS-CNAM, PARIS

German cabinet-maker David Roentgen's eighteenth-century automaton of Marie Antoinette plays a table-top zither.

TECHNOLOGY

The automaton chronicles

Our responses to robots echo down the millennia, say **Stephen Cave** and **Kanta Dihal**.

The French philosopher René Descartes was reputedly fond of automata: they inspired his view that living things were biological machines that function like clockwork. Less known is a strange story that began to circulate after the philosopher's death in 1650. This centred on Descartes's daughter Francine, who died of scarlet fever at the age of five.

According to the tale, a distraught Descartes had a clockwork Francine made: a walking, talking simulacrum. When Queen Christina invited the philosopher to Sweden in 1649, he sailed with the automaton concealed in a casket. Suspicious sailors forced the trunk open; when the mechanical child sat up to greet them, the horrified crew threw it overboard.

The story is probably apocryphal. But it sums up the hopes and fears that have been associated with human-like machines for nearly three millennia. Those who build such devices do so in the hope that they will overcome natural limits — in Descartes's case, death itself. But this very unnaturalness terrifies and repulses others. In our era of advanced robotics and artificial intelligence (AI), those polarized responses persist, with pundits and the public applauding or warning against each advance. Digging into the deep history of intelligent machines, both real and imagined, we see how these attitudes evolved: from fantasies of trusty mechanical helpers to fears that runaway advances in technology might lead to creatures that supersede humanity itself.

Arguably the oldest known story of something approximating AI can be found in the eighth-century-BC *Iliad*, Homer's epic poem of the Trojan War. In it, Hephaestus, disabled god of metalworking, creates golden handmaidens to help him in his forge: "In them is understanding in their hearts, and in them speech and strength, and they know cunning handiwork". Hephaestus was ▶

© 2018 Springer Nature Limited. All rights reserved.

also supposedly responsible for the first 'killer robot', Talos. A mechanical bronze colossus featuring in the third-century-BC epic *Argonautica*, it patrolled the shores of Crete, lobbing boulders at invaders.

These fictions were grounded in reality: ancient Greek technologists were astonishingly skilled in mechanics and metalwork. In her forthcoming book *Gods and Robots*, classicist Adrienne Mayor describes bronze automata that featured at the Olympic Games two centuries before the *Argonautica* — a leaping dolphin and eagle in apparent flight. In his first-century-AD treatise *On Automaton-Making*, the mathematician-engineer Hero of Alexandria describes a fully automated puppet theatre that, through a combination of displaced grain, axles, levers, pulleys and wheels, could enact an entire tragedy.

These classical stories reveal how, then as now, humanoid machines were mostly conceived as representing straightforward hopes — the ideal servant who always obeys, the perfect soldier who never tires. But as the influence of Greece declined over the first centuries AD, the West entered a millennium in which the skills of automaton-making were lost, along with the aspirations associated with them. As historian E. R. Truitt describes in *Medieval Robots* (2015), it was the Byzantine Empire and Arab world that preserved the mechanical arts over those centuries. Around AD 850, for instance, the Banu Musa brothers in what is now Iraq published the *Book of Ingenious Devices*, which featured automata such as a water-powered organ. Thus, in the West, the 'otherness' of mechanical simulacra was compounded. Associated with the exotic and the idea of an 'infidel' East, automata were viewed for some time with awe and suspicion.

The thirteenth century saw a resurgence of Western interest. Automata began appearing at courts as showpieces designed to wow visitors. The Count of Artois at the château of Hesdin in what is now northern France commissioned a cluster of mechanized beasts and androids that interacted with guests, scolding them and or even soaking them with water. At the same time, hints of darker themes began to emerge. A number of great medieval scholars, such as Roger Bacon and Albertus Magnus, were each rumoured to have created a bronze head that could answer any question — a proto-Siri. These stories end badly, with the oracle's destruction, sometimes by people who mistrust it. In such cautionary tales, the creation of AI is an act of Promethean hubris, semi-divine power no mortal should possess — presaging the story of Descartes's daughter and Mary Shelley's *Frankenstein* (see R. Holmes *Nature* **535**, 490–492; 2016).

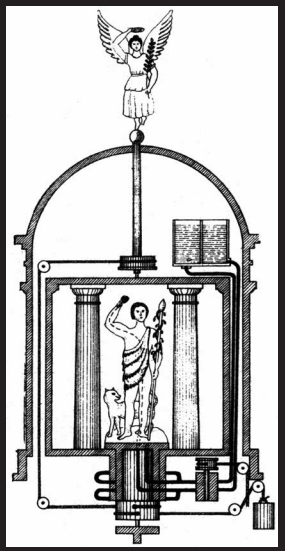
The march of mechanical humanoids nevertheless persisted into the great cultural and technological flowering of the Renaissance. Evoking widespread wonder, hydraulic, spring-powered and clockwork automata proliferated in Europe, from mechanized angels in churches to Neptune automata in grottoes. Their mystical connotations began to give way to reflections of cutting-edge scientific developments: Leonardo da Vinci, for instance, drew up plans for a robotic knight operated internally by weights and pulleys.

Over subsequent centuries, Descartes's view of living things as complex machines inspired new heights in manufacture, and Europe reached peak automaton in the seventeenth to early nineteenth centuries. Master artisans built marvels of art-imitating-life, such as Jacques de Vaucanson's 1739 Digesting Duck. Sporting more than 400 moving parts in each wing and internal rubber tubing, it appeared to eat, drink and defecate.

But the duck had no 'digestive' mechanism: its 'droppings' were actually prefabricated pellets ejected from a hidden compartment. A few decades later in 1770, inventor Wolfgang von Kempelen's infamous chess-playing Mechanical Turk debuted. Although many suspected it was a hoax, nobody could quite work out how they were fooled until a revelatory article in 1857 announced that it was operated by a person hidden inside. Awe over the transgressive nature of such mechanisms began to mingle with fears of deception, as reflected in the fiction of the time. In Prussian Romantic author E. T. A. Hoffmann's 1816 short story 'The Sandman', for example, the protagonist Nathanael is bewitched by the beauty of a woman called Olimpia. The discovery that she is an automaton drives him to suicide.

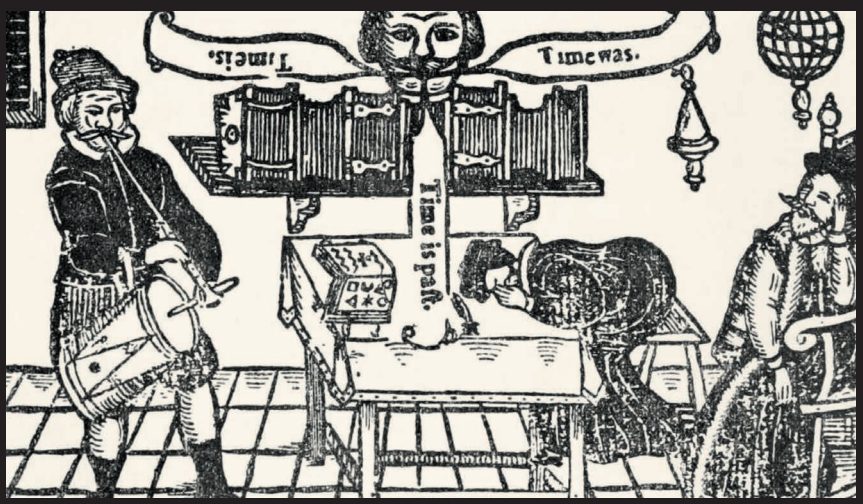
The imaginative resonance of intelligent machines began to reach its fullest in the twentieth century, when industrialization

Below: A nineteenth-century engraving of an ancient Greek puppet theatre.





Above: The eighteenth-century Mechanical Turk.



A 1630 woodcut depicts the story of scholar Roger Bacon's talking bronze head.

TOP LEFT: INTERFOTO/ALAMY; TOP RIGHT: INTERFOTO/ALAMY; BOTTOM: ANON./WIKIMEDIA COMMONS

© 2018 Springer Nature Limited. All rights reserved.

replaced the rhythms of nature with those of the production line. It was also a time of revolution and mechanized warfare. Against that backdrop, the word 'robot' was born in Czech writer Karel Čapek's 1920 play *R.U.R. (Rossum's Universal Robots)*. In the very work coining the term, the robots rebel against and destroy their creators. And that narrative of rebellion has proved to be the most potent of all our AI fears, retold repeatedly as technology evolves.

During the cold war space race, the film *2001: A Space Odyssey* (1968) gave us HAL 9000, the murderous spaceship supercomputer. With the rise of the Internet, we got Skynet — a defence network that becomes self-aware in the Terminator films (starting in 1984) — and *The Matrix* (1999), featuring intelligent machines that farm humans whose minds unknowingly inhabit a virtual reality. Now, with AI dominating headlines, we have sophisticated robots again overthrowing their wetware masters, from Ava in the 2015 film *Ex Machina* to the android amusement-park hosts in the *Westworld* television series.

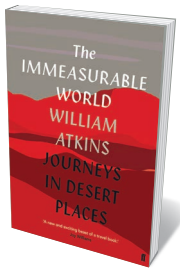
"The tension lies in our conflicted desire to create beings superhuman in capacity, but subhuman in status."

The persistent trope of robot revolts reveals the paradox at the heart of our relationship with intelligent machines. We want to create clever tools that can do everything we can do, and more. They will be the perfect oracles, servants, soldiers, even lovers. To fulfil our hopes, they must have attributes such as intellect and agency — minds of their own, superior to ours. But, paradoxically, that is also why we fear HAL and Skynet. The tension lies in our conflicted desire to create beings superhuman in capacity, but subhuman in status.

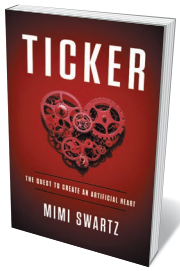
Our hopes continually threaten to collapse into such fears, but we hope nevertheless. Every robot rebel has its benevolent counterpart, such as C-3PO in the Star Wars franchise or the android child David in Steven Spielberg's 2001 film *A.I. Artificial Intelligence*. Both kinds of stories, the hopeful and the fearful, reveal to us our complex emotional responses to AI. Understanding these and their deep history is crucial to making the most of life with intelligent machines. ■

Stephen Cave *is executive director of the Leverhulme Centre for the Future of Intelligence (CFI) at the University of Cambridge, UK.* **Kanta Dihal** *is the postdoctoral researcher on the AI Narratives project at CFI. They are co-authors of the forthcoming book* AI: A Mythology. *e-mails: sjc53@cam.ac.uk; ksd38@cam.ac.uk*

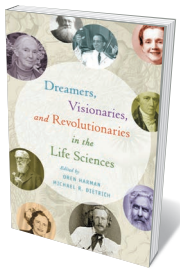
Books in brief



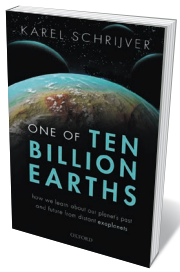
The Immeasurable World
William Atkins FABER & FABER *(2018)*
An "axis where the absolute coexists with the infinite": so William Atkins sees the quintessence of deserts. He searched for it on seven journeys, from the Arabian Peninsula's 'Empty Quarter' to China's Gobi Desert, recollecting his arid odyssey in this beautifully observed memoir-cum-natural-history. With him, we tread the grit of old atomic test sites in Australia's Nullarbor Plain; meet fatalistic migrants at Arizona's Mexican border; and contemplate the bleak ecodisaster of Central Asia's disappearing Aral Sea. "The world has been done," Atkins declares. But he sees it with new eyes.



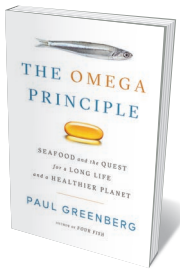
Ticker: The Quest to Create an Artificial Heart
Mimi Swartz CROWN *(2018)*
In this pacy, blow-by-blow account of the search for a viable artificial heart, *Texas Monthly* executive editor Mimi Swartz offers a vivid gallery of the medical pioneers who have jostled for the prize. Here are heart surgeon O. H. 'Bud' Frazier; inventors Robert Jarvik, Billy Cohn and Daniel Timms; and cardiac specialists Michael DeBakey and Denton Cooley, whose duel over implanting the first crude devices sparked the crusade. Decades later, that quest remains urgent: more than 25 million US citizens have heart disease, yet only a few thousand hearts are available for transplant in any given year.



Dreamers, Visionaries, and Revolutionaries in the Life Sciences
Oren Harman and Michael R. Dietrich (eds) UNIV. CHICAGO PRESS *(2018)*
They may infuriate even as they inspire; destabilize, as well as advance. The visionaries of biology, note science historians Oren Harman and Michael Dietrich, are masters of "dramatically bold, even fantastic" thinking about big problems. This compelling edited volume explores the work of 19 innovators, including Iranian-American cancer researcher Mina Bissell, who studies tumour microenvironments; Canadian John Todd, who engineered solar aquatic sewage treatment; and Peter Kropotkin, the Russian biologist who championed mutualism in nature during the Darwinian revolution.



One of Ten Billion Earths
Karel Schrijver OXFORD UNIV. PRESS *(2018)*
Like distant mirrors, exoplanets reveal much about Earth and its habitability even as they relay explosive new insights into planetary formation and extraterrestrial life. Astrophysicist Karel Schrijver's lucid, fact-packed primer ranges over everything from the Goldilocks zone and stellar nurseries to disrupted exosystems and the vagaries of "living on a pale blue dot". And more: if you ever wanted to understand nomad planets or catch a glimpse of a volcanic eruption on Jupiter's moon Io (observed by the New Horizons spacecraft on its way to Pluto), this book delivers the goods.



The Omega Principle
Paul Greenberg PENGUIN *(2018)*
The cacophony of claims about the impact of omega-3 fatty acids on health is drowning out concerns over the associated industry, argues Paul Greenberg in this sobering investigation. He followed the omega-3 trail around the world and into history to reveal the price of a century of industrial reduction — the boiling down of a gargantuan tonnage of marine organisms for products from animal feed to oils for supplements. Notably, the mass harvesting of krill and forage fish such as menhaden looks increasingly unsustainable as overfishing and oceanic warming take their toll. **Barbara Kiser**

© 2018 Springer Nature Limited. All rights reserved.

Correspondence

Earth-observation data must stay free

US President Donald Trump's administration is considering whether to charge for access to Earth-observation data (*Nature* **556**, 417–418; 2018). As former director of the Group on Earth Observations (GEO), I urge the US government not to falter in its global leadership in open-data policies or to dismiss the importance of their worldwide benefits, particularly for the United States.

As you point out, this would not be the first time that the United States has charged for Landsat satellite data. Until 2008, it used a cost-of-fulfilling-user-requests model, bringing in about US$4.5 million annually. The largest purchaser of data was the federal government itself, so its net profit was negligible. Meanwhile, external research and innovation withered.

Once replaced by an open-access policy, economic returns followed: in 2011, for example, these were $1.7 billion for the United States and $400 million for other countries, bringing the global total to $2.1 billion. Today, there are more than 400 million international open Earth-observation data and information resources (see www.geoportal.org).

We must remain committed to keeping Earth-observation data — collected at taxpayers' expense — freely and readily available to inform policy decisions, including those called for in the Sendai Framework for Disaster Risk Reduction, the United Nations 2030 Agenda for Sustainable Development and the Paris climate agreement.

Barbara J. Ryan *Bonita Springs, Florida, USA.*
bjrgeneva@gmail.com

Predict trajectory, not emergence

It is important to distinguish between predicting the emergence of infectious diseases and their trajectory (see E. C. Holmes *et al. Nature* **558**, 180–182; 2018). It is currently impossible to predict emergence, but feasible and effective to predict the trajectory.

For example, modelling of established outbreaks was important for responding to those caused by the Ebola, Zika, H1N1 influenza and severe acute respiratory syndrome (SARS) viruses. It helped decision-makers to plan bed capacity (M. I. Meltzer *et al. MMWR Morb. Mortal. Wkly Rep.* **65**, 85–89; 2016), anticipate travel-related spread (I. I. Bogoch *et al. Lancet* **385**, 29–35; 2015) and plan vaccine trials (A. Camacho *et al. Vaccine* **35**, 544–551; 2017).

Several US government agencies maintain such capabilities, including the Centers for Disease Control and Prevention, and the Biomedical Advanced Research and Development Authority. And the White House Office of Science and Technology Policy published a report in 2016 that identified epidemic prediction and forecasting as a crucial tool in outbreak preparedness and response (see go.nature.com/2ubIwjt).

Caitlin M. Rivers *Johns Hopkins Center for Health Security, Baltimore, Maryland, USA.*
Samuel V. Scarpino *Network Science Institute, Northeastern University, Boston, Massachusetts, USA.*
crivers6@jhu.edu

Biomedical success on the cheap

A handful of science students at Bellevue College, a small community institution in the United States, have pioneered their own training in biomedical research methods — with no wet lab facilities and on a minimal budget.

Our group identified a promising project by tapping into advances in computational analysis and the vast, readily accessible amounts of genomics data. We decided to search for viral protein sequences that could be important for tissue tropism.

Human pathogenic viruses were classified according to the tissue they infect (pulmonary, gastrointestinal, and so on) — irrespective of whether they were enveloped or non-enveloped RNA or DNA viruses. The students majoring in computer science developed programs to speed up protein-sequence comparisons and alignments in each category. The biologists identified consensus protein domains.

Within a month, the students had localized a consensus amino-acid sequence on the adherence spikes of the chikungunya virus (CHIKV). A literature search revealed that this sequence is responsible for binding to a receptor on human cells and infecting them (R. Zhang *et al. Nature* **557**, 570–574; 2018). They then came up with the idea that this domain could be used as an antigen for producing a vaccine against CHIKV. Clinical trials of virus-like particles are ongoing (see go.nature.com/2nzgvpw and go.nature.com/2lhwjrh).

It is heartening that a group of young scientists in a community institution, who had never before heard of CHIKV, achieved so much as a result of their curiosity and motivation.

Reza Forough* *Bellevue College, Bellevue, Washington, USA.*
reza.forough@bellevuecollege.edu
**On behalf of 4 correspondents; see go.nature.com/2ubuchn*

Health-care impacts of web shutdowns

Governments in Africa, Asia and the Middle East are increasingly shutting down the Internet as a blunt response to threatening situations such as political unrest (see go.nature.com/2msruv2). We call on inter- and non-governmental organizations to put pressure on governments to evaluate the potentially devastating effects of this tactic on health care.

In Cameroon, for example, the government cut off the Internet for three months last year. This disrupted crucial services such as the GiftedMom smartphone app, which connects rural mothers with physicians and guides them through infant vaccinations, in a country where children are almost twice as likely as the global average to die before their fifth birthday.

Furthermore, Cameroon was unable to submit 85% of its health-performance data to the District Health Information System data set DHIS2, which is used to guide funding decisions. Health workers received only partial pay because they could not report their progress to the World Bank's performance-based financing system.

Yet no empirical research exists for the health impacts of such shutdowns. We need to know, for example, the risks they pose to emergency services and to funding for Internet-enabled medicine. Policymakers must have comprehensive data on interruptions to online services if health outcomes in the developing world are to be improved.

Okwen Mbah, **Miriam Nkangu** *Effective Basic Services (eBase) Africa, Nkwen Bamenda, Cameroon.* **Zak Rogoff** *Access Now, Washington DC, USA.*
okwen@ebaseafrica.org

CONTRIBUTIONS
Correspondence may be submitted to **correspondence@nature.com** after consulting the author guidelines and section policies at **http://go.nature.com/cmchno**.

© 2018 Springer Nature Limited. All rights reserved.

ARCHAEOLOGY

An early hominin arrival in Asia

An excavation has uncovered stone tools in China that are about 2.1 million years old. This evidence pushes back the date of the earliest established signs of a hominin species outside Africa. SEE LETTER P.608

JOHN KAPPELMAN

Imagine a world nearly devoid of humans. This theme is often explored in post-apocalyptic science fiction, and it was the case for most of Earth's history until early hominins (members of the evolutionary tree that includes humans, extinct species of the genus *Homo* and other closely related bipedal species) moved out of Africa into unknown territories, to eventually populate the planet. Which hominins made this journey? When and how did they migrate, and by which routes? Field-based research that uncovers traces of these early dispersals could provide some answers, as well as insight into hominin behaviour. On page 608, Zhu *et al.*[1] report evidence of hominin activity in China more than 2 million years ago, revealing an earlier time frame than was previously known for a hominin presence outside Africa.

Establishing the first known occurrence of a species requires incontrovertible evidence to confirm the presence of the species and firm support for the age of the geological material that contains such a sample. The remains of species that were common where they lived are usually found with ease in fossil deposits. By contrast, populations of ancient hominins were probably rare, and their fossilized remains are generally scarce. Just a single finger bone can suffice to document a hominin presence[2]. However, when hominins began to fashion tools from stones, at least 3.3 million years ago[3], the purposely chipped cobbles and flakes that they produced became another calling card that can attest to a hominin presence.

Until now, the oldest known hominin site outside Africa[4] was in Dmanisi, Georgia. Excavations at that site uncovered spectacular finds of the roughly 1.85 million- to 1.78-million-year-old remains of multiple hominins and stone tools. Numerous later sites of hominin activity, at locations stretching from western Europe[5] to eastern Asia[6], have also been investigated thoroughly. Zhu and colleagues' report of signs of a hominin presence at Shangchen in China's Loess Plateau (Fig. 1) is based on evidence from only stone tools, and the researchers found that these tools were distributed in layers of sediment that date back to about 2.1 million years ago.

The age of many hominin sites has been estimated by methods such as radiometric dating, or by the chemical fingerprinting of volcanic rocks. Because Shangchen lacks volcanic rocks, Zhu and colleagues instead used palaeomagnetic dating to analyse sediment layers called palaeosols (fossilized soils) and loess (wind-blown silts). This technique relies on the fact that Earth's magnetic field undergoes random reversals[7], in which the magnetic north pole becomes the magnetic south, and vice versa. Magnetic minerals in sediments act like small compasses that register the polarity, and when such sediments become rock, the polarity of that time is locked in. The pattern of polarity reversals in these ancient sediment layers provides a fingerprint that can be matched to a dated reference called the geomagnetic polarity timescale (GPTS)[8].

The site at Shangchen contains steep, deeply incised outcrops that include slopes covered in vegetation, which makes sample collection a literal cliffhanger. The authors' data provide a convincing correlation to the GPTS, and the team was able to identify two notable events from the GPTS (the brief Réunion Subchron[9] and the later, but slightly longer, Olduvai Subchron), which bookend the layer that contains the oldest artefacts. Can we be certain that the researchers found the oldest tools at the site? The team's ability to reach the deeper layers was limited because of the present active farming of the fields. Investigating such layers should be a goal of future work.

The 96 excavated stone tools that Zhu *et al.* analysed were mainly small stone flakes and cobbles, which contrasted sharply with the fine-grained material that surrounded the artefacts. The authors propose that hominins transported these rocks from the Qinling Mountains to the south. If the exact source of the stones could be determined, it would provide important information about how far these hominins transported raw materials.

The stone tools are simple in form: the palm-sized rock cores had a limited number of flakes removed, with some of the flakes showing an apparent subsequent resharpening, and several cobbles might have been used as hammerstones. These characteristics closely align the Shangchen tools with those of a similar age found in Africa[10]. The authors do not report any matching refit of flakes to

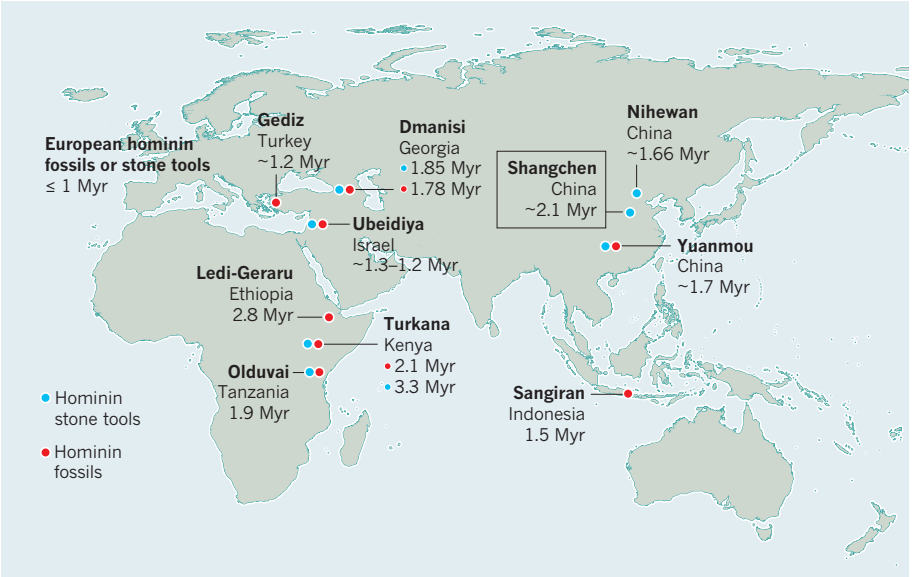


Figure 1 | Ancient sites of hominin presence. Zhu *et al.*[1] report their discovery of approximately 2.1-million-year-old stone tools at Shangchen in China's Loess Plateau, which provides the earliest known evidence for the presence of hominins (the evolutionary group that includes humans, extinct species of the genus *Homo* and related bipedal species) outside Africa. The dates (Myr, million years ago) and locations of some of the important earliest known sites of hominin fossils and stone tools are shown.

© 2018 Springer Nature Limited. All rights reserved.

each other or to the cores from which they were chipped, so flake production might have occurred elsewhere. However, because the steepness of the slopes at the site rarely permitted the team to open large excavations, this prospect remains to be examined thoroughly.

What were the stone tools used for? Zhu *et al.* note the discovery of animal remains near the oldest tools, including bones belonging to bovids (a family that includes antelopes and cattle), cervids (comprises deer) and suids (pigs). The authors did not address whether this association provides evidence of tool use for carcass processing. To evaluate the possibility, it would be necessary to identify signs such as: cut marks on the bones that point to flesh removal with tools; breakage marks on the bones, suggesting that they had been hammered to extract bone marrow; tool wear; or the presence of trace biological residues on the tools. If this question is investigated in the future, the degree of post-recovery cleaning of the stone tools might have compromised the use of approaches[11] such as residue studies.

Hominins originated in Africa possibly more than 6 million years ago[12]. The absence in Eurasia of both any hominin sites dating to the early portion of this interval and any fossils that can be attributed to hominin genera such as *Australopithecus* and *Paranthropus*, found in Africa until about 1 million years ago, points instead to a species of *Homo* as the most likely candidate for the first hominin to have left Africa. The oldest known African fossil attributed to *Homo* is a 2.8-million-year-old jawbone from Ethiopia[13], which provides a time estimate for the earliest possible exit of the genus from Africa. Of course, the actual date of departure might have been later.

The hominin dispersal probably occurred under the variable climates of the Pleistocene ice age. Does a migration to higher latitudes suggest the evolution of behavioural adaptations to colder climates? Perhaps. The conventional interpretation that the palaeosols of the Loess Plateau formed during favourable warm and wet conditions, and its loess under harsher cold and dry conditions, is probably an oversimplification[14], but at the Shangchen site, palaeosol layers containing stone tools outnumber loess layers containing such tools by a ratio of about 2 : 1. Rather than maintaining a continuous occupation of the Loess Plateau, the hominin population might have increased or dwindled, depending on the climate[15].

The roughly 14,000-kilometre trek from eastern Africa to eastern Asia represents a range expansion of dramatic proportions. The dispersal of hominins was probably facilitated by population increases as they moved into new territories and filled empty niches, and could also have been driven by the phenomenon of resource depletion that underlies the high mobility of today's hunter-gatherers[16]. Yet even with a dispersal rate of only 5–15 kilometres per year, a value well inside the daily foraging range of modern hunter-gatherers[17], the distance between Africa and Asia could have been covered in just 1,000–3,000 years. The present record of hominin sites and the dating techniques that are currently available to researchers are not sufficient to resolve a dispersal event of such potential speed, or to determine its exact form, but we can surely look forward to more finds that will help to solve this migration mystery. ■

John Kappelman *is in the Departments of Anthropology and Geological Sciences, The University of Texas at Austin, Austin, Texas 78712, USA.*
e-mail: jkappelman@austin.utexas.edu

1. Zhu, Z. *et al. Nature* **559**, 608–612 (2018).
2. Groucutt, H. S. *et –al. Nature Ecol. Evol.* **2**, 800–809 (2018).
3. Harmand, S. *et al. Nature* **521**, 310–315 (2015).
4. Ferring, R. *et al. Proc. Natl Acad. Sci. USA* **108**, 10432–10436 (2011).
5. Muttoni, G., Scardia, G. & Kent, D. V. *Quat. Sci. Rev.* **180**, 1–29 (2018).
6. Zhu, R. X. *et al. Nature* **431**, 559–562 (2004).
7. Phillips, J. D. & Cox, A. *Geophys. J. Int.* **45**, 19–33 (1976).
8. Ogg, J. G. in *The Geologic Time Scale 2012* (eds Gradstein, F. M., Ogg, J. G., Schmitz, M. D. & Ogg, G. M.) 85–114 (Elsevier, 2012).
9. Quidelleur, X., Holt, J. W., Salvany, T. & Bouquerel, H. *Geophys. J. Int.* **182**, 699–710 (2010)
10. Shea, J. J. in *Out of Africa I* (eds Fleagle, J. G., Shea, J. J., Grine, F. E., Baden, A. L. & Leakey, R. E.) 47–64 (Springer, 2010).
11. Nowell, A. *et al. J. Archaeol. Sci.* **73**, 36–44 (2016).
12. Klein, R. G. *The Human Career* 3rd edn (Univ. Chicago Press, 2009).
13. Villmoare, B. *et al. Science* **347**, 1352–1355 (2015).
14. Stevens, T. *et al. Nature Commun.* **9**, 983 (2018).
15. Prat, S. *C. R. Palevol* **17**, 6–16 (2018).
16. Venkataraman, V. V., Kraft, T. S., Dominy, N. J. & Endicott, K. M. *Proc. Natl Acad. Sci. USA* **114**, 3097–3102 (2017).
17. Kelly, R. L. *The Lifeways of Hunter-Gatherers* (Cambridge Univ. Press, 2013).

This article was published online on 11 July 2018.

QUANTUM PHYSICS

Quantum optics without photons

Atoms can exhibit wave-like behaviour to form matter waves. Such waves have been used to model the basic processes that underpin how light interacts with matter, providing an experimental platform for future research. SEE LETTER P.589

ALEJANDRO GONZÁLEZ TUDELA & J. IGNACIO CIRAC

The fundamental theory that describes the interaction between light and matter at microscopic scales is known as quantum optics. One of the most striking and tangible consequences of this theory is that excited atoms in the quantum electrodynamic vacuum — a state that contains no photons, often referred to simply as the vacuum — can decay to their ground state by emitting a photon, a process called spontaneous emission. In 1946, physicist Edward Mills Purcell proposed that this process can be tailored by structures that alter the photonic environment[1], such as photonic crystals[2], engineered dielectrics (insulators) in which light cannot propagate at certain frequency ranges. In the 1990s, it was predicted[3] that spontaneous emission in photonic crystals leads to exotic decays, in which photons

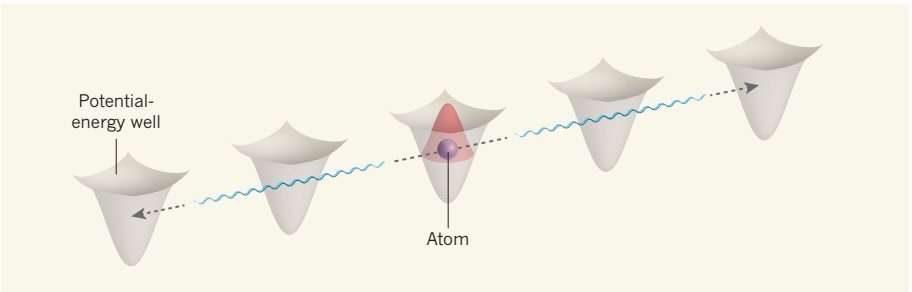


Figure 1 | Matter-wave emitters. Krinner *et al.*[4] trapped rubidium atoms in a one-dimensional optical lattice formed by the interference of laser fields — the interference generates a periodic pattern of light intensity that corresponds to a series of potential-energy wells, in which atoms can be confined. Each atom has two internal states, one of which (red) is confined to the atom and the other (blue) which is unconfined and overlaps with adjacent empty wells. By tuning the parameters of the system, the authors could make occupied lattice sites emit atoms as waves (dotted arrows) that travel along the lattice. The emission mimics that of photons from atoms confined in photonic crystals (materials engineered so that light cannot propagate at certain frequency ranges).

© 2018 Springer Nature Limited. All rights reserved.

adopt a 'superposition' state in which they are simultaneously emitted into the surrounding environment but also localized around the emitting atoms. On page 589, Krinner *et al.*[4] report the first observation of the dynamics of these exotic decays, not using photons, but using a system of trapped, ultracold atoms.

The atoms in Krinner and colleagues' experiments are trapped in optical lattices[5], which are formed by the interference of counter-propagating laser fields — the interference generates a periodic pattern of light intensity in which the atoms are confined. Given their quantum-mechanical nature, the atoms can tunnel between neighbouring sites in the lattice at a rate that can be adjusted by altering the lasers' intensities. Because such systems are highly controllable and exhibit low decoherence (the atoms are well isolated from their environment and thus behave ideally), they are an almost perfect platform for simulating complex quantum problems found in fields such as condensed-matter and high-energy physics (see go.nature.com/2uied19). Krinner *et al.* now demonstrate that trapped-atom systems can also be used to simulate quantum optical problems.

The authors' experiments are based on a proposal[6] published in 2008. The idea is to use atoms — in this case, rubidium atoms — that have two internal states, which respond to a one-dimensional optical lattice in different ways (Fig. 1). One state (let's call it the *f* state) 'sees' regions of high light intensity as deep optical potential-energy wells from which it cannot move, whereas the other (the *a* state) hardly notices the wells, so that atoms in that state can propagate through the optical lattice as a matter wave. An atom in state *f* thus represents a matter-wave emitter in an excited state, whereas an atom in state *a* behaves like a photon that can be emitted through spontaneous emission. To complete the analogy with atomic decay phenomena in a structured photonic environment, the two internal states are coupled to each other using external fields, so that an initial excitation can be transformed into a propagating matter wave.

In their experiments, Krinner *et al.* model the simplest scenario of spontaneous emission in a photonic crystal, but one that potentially offers the most insight into such processes: the spontaneous emission of a single photon into the vacuum. By tuning the experimental parameters of their system, they observed a phenomenon known as fractional decay[5], in which the emitter ends up in a quantum superposition of being both excited and having decayed to the ground state. The authors also report direct evidence that the probability of the emitter remaining in the excited state does not decrease exponentially over time. Both the non-exponential behaviour and the fractional decay are some of the most peculiar effects that photonic crystals induce in quantum emitters. The new measurements are analogous to previously reported measurements of the spontaneous emission of photons in the visible-light[7,8] and microwave[9] regions of the electromagnetic spectrum, but in those studies, it was not possible to measure the dynamics of the decay.

The authors' experimental platform offers several useful features in addition to its excellent controllability and low decoherence. First, it is flexible enough to emulate photonic crystals that have different geometries, and to model 3D environments in which further unusual features of spontaneous emission emerge[10,11]. It also allows access to parameter regimes that are out of reach of optical implementations, such as situations in which the coupling between emitters and the environment is very strong. Moreover, it might enable spontaneous emission to be studied in environments that are even more exotic than photonic crystals, such as in materials known as topological insulators — although the experimental set-up would need to be adapted so that the atoms move in the way that simulates the movement of excitations in such materials, and the efficiency with which decays can be detected would need to be improved.

Together with photonic platforms in the optical and microwave regimes, Krinner and colleagues' system opens the way to studies of the physics that emerges in unconventional quantum optical set-ups. If the experiments are extended to include many emitters, it might be possible to observe collective spontaneous-emission phenomena that cannot be predicted using current computational methods, or even to engineer interactions among the emitters that cannot be produced using other platforms. ■

Alejandro González Tudela *and* **J. Ignacio Cirac** *are in the Theory Division of the Max Planck Institute of Quantum Optics, Garching, D-85748 Germany.*
e-mail: ignacio.cirac@mpq.mpg.de

1. Purcell, E. M. *Phys. Rev.* **69**, 681 (1946).
2. Yablonovitch, E. *Phys. Rev. Lett.* **58**, 2059 (1987).
3. John, S. & Quang, T. *Phys. Rev. A* **50**, 1764 (1994).
4. Krinner, L., Stewart, M., Pazmiño, A., Kwon, J. & Schneble, D. *Nature* **559**, 589–592 (2018).
5. Bloch, I., Dalibard, J. & Zwerger, W. *Rev. Mod. Phys.* **80**, 885 (2008).
6. De Vega, I., Porras, D. & Cirac, J. I. *Phys. Rev. Lett.* **101**, 260404 (2008).
7. Hood, J. D. *et al. Proc. Natl Acad. Sci. USA* **113**, 10507 (2016).
8. Hoeppe, U. *et al. Phys. Rev. Lett.* **108**, 043603 (2012).
9. Liu, Y. & Houck, A. A. *Nature Phys.* **13**, 48–52 (2017).
10. González-Tudela, A. & Cirac, J. I. *Phys. Rev. Lett.* **119**, 143602 (2017).
11. González-Tudela, A. & Cirac, J. I. *Phys. Rev. A* **97**, 043831 (2018).

MICROBIOLOGY

Viruses cooperate to defeat bacteria

It emerges that viruses called phages, which infect bacteria, can suppress the bacterial immune system during an initial wave of unsuccessful infection, enabling subsequent viral infection to succeed.

AUDE BERNHEIM & ROTEM SOREK

Bacteria and the viruses that infect them, known as phages, are engaged in a constant arms race. Bacteria continually evolve new mechanisms of resistance against viruses, while phages evolve countermeasures to overcome these defence mechanisms. Writing in *Cell*, Borges *et al.*[1] and Landsberger *et al.*[2] reveal how phages can 'collaborate' to shut down the bacterial immune system and achieve a successful infection. Although an initial viral attempt at infection fails, this enables a subsequent phage infection to be successful. This example of 'cooperation' between genetically identical individuals of a viral population illuminates a previously unknown group strategy of phages, and provides an interesting example of viral 'altruism'.

The bacterial anti-phage defence system called CRISPR–Cas recognizes and targets foreign nucleic acids in a sequence-specific manner[3]. To block CRISPR–Cas defences, phages express genes that encode proteins that inhibit the function of the CRISPR–Cas machinery[4,5]. Because these anti-CRISPR proteins are encoded in the phage genome (Fig. 1), a phage must enter a bacterial host cell and begin to express them to mount a counter-attack. Yet this raises a conundrum. CRISPR–Cas defences can attack phage DNA as soon as it enters a bacterial cell[6], before the phage gets a chance to express and use its anti-CRISPR proteins. So what purpose do anti-CRISPR proteins serve if the help they provide is likely to arrive too late for the individual virus that expresses them?

To investigate this, Borges, Landsberger and their respective colleagues studied the bacterium *Pseudomonas aeruginosa* and examined viral anti-CRISPR proteins that target a version of the bacterial CRISPR–Cas system called type I-F. Both groups observed that the initial ratio between the number of phages and the

© 2018 Springer Nature Limited. All rights reserved.

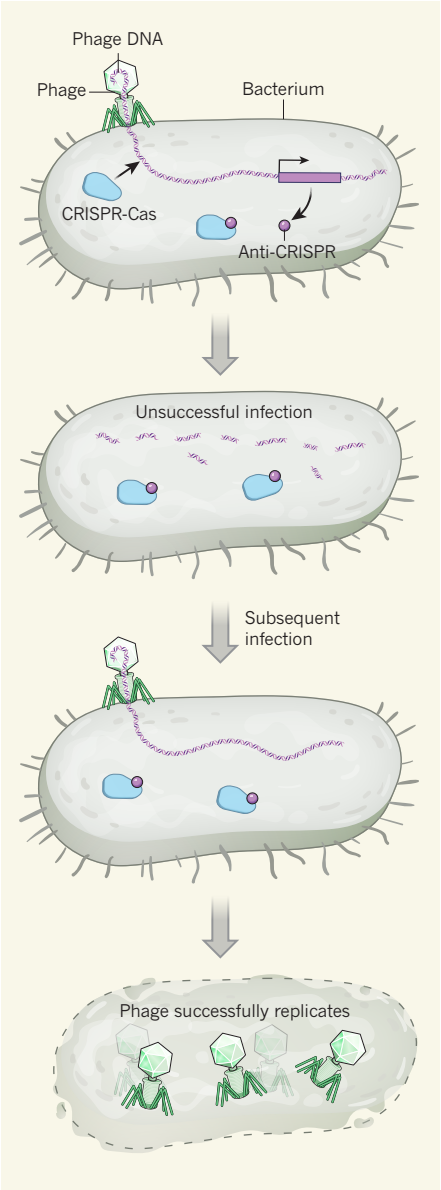


Figure 1 | A viral mechanism to thwart bacterial defences. The bacterial protein and RNA complex termed CRISPR–Cas recognizes and destroys foreign DNA, such as that of invading bacteria-infecting viruses called phages, in a sequence-specific manner[3]. Some phages encode anti-CRISPR proteins that bind to CRISPR–Cas and inhibit its function[4,5]. However, how such proteins aid viruses has been unclear, given that an initial infecting phage is probably targeted and destroyed by CRISPR–Cas, even if the phage expresses anti-CRISPR proteins before it is eliminated. Borges *et al.*[1] and Landsberger *et al.*[2] provide evidence that an initial unsuccessful phage infection that leads to expression of anti-CRISPR proteins generates an immunosuppressed bacterium that could then succumb to a subsequent phage infection, thereby facilitating successful phage replication.

number of bacterial host cells, known as the multiplicity of infection (MOI), could be used to predict whether phage infection would be successful. When the authors tested bacteria that lacked CRISPR–Cas immunity, even a low dose of phages resulted in successful viral propagation that caused the eventual demise of the bacterial culture. However, for bacteria and phage populations encoding CRISPRs and anti-CRISPRs, respectively, phages managed to replicate only if the initial MOI was high.

The two groups followed different strategies to determine why the initial ratio of phages to bacteria is important for the success of phage propagation. Borges and colleagues used a genetic approach to test whether the inactivation of bacterial defences was due to the presence of anti-CRISPR proteins from more than one phage in the same host cell. The authors engineered a phage that was incapable of replication but did express anti-CRISPR proteins. By mixing these engineered phages with wild-type ones, the authors showed that the presence of the engineered phages could enable a wild-type phage infection to succeed even at a low MOI that would normally fail.

Landsberger and colleagues, by contrast, created a mathematical model to investigate the observed MOI dependence of phage infections. They could recapitulate their experimental findings only by using a model simulation in which bacteria enter an immunosuppressed state after surviving an initial infection by a phage that expressed anti-CRISPR proteins.

Both groups' results suggest that the first wave of phages that attempt infection succumb to CRISPR–Cas defences but manage to deliver anti-CRISPR proteins that immunosuppress the bacterial cell. This initial attack paves the way for a second wave of phages to successfully infect the now-defenceless bacterium. Having a high MOI increases the probability that a second wave of phage infection will occur.

The authors of these two studies validated this hypothesis in different ways. Landsberger and colleagues infected a bacterial population with phages encoding anti-CRISPR proteins and then introduced a type of circular DNA called a plasmid into the bacteria. Plasmid DNA would normally be targeted by the bacterial immune system, but the earlier phage infection limited this targeting of plasmid DNA, demonstrating that the bacterial cells were in an immunosuppressed state. Borges *et al.* demonstrated that phage-mediated immune suppression also occurs for another type of CRISPR–Cas system, called type II, suggesting that phage cooperation to enable immunosuppression of bacteria is a general principle that extends beyond a single type of CRISPR–Cas system.

The studies by Borges, Landsberger and their colleagues provide fresh insights into viral group dynamics, and join a growing body of evidence indicating that viruses can benefit from the group behaviour of a viral population. For example, phages can coordinate their infection dynamics using communication by small molecules to determine whether an individual phage will replicate in an infected cell or enter a dormant state termed lysogeny[7,8]. Perhaps phages also cooperate to overcome

© 2018 Springer Nature Limited. All rights reserved.

components of the bacterial immune system other than CRISPR–Cas, such as restriction enzymes or other antiviral defence systems[9].

Many questions remain to be answered concerning the bacterial immunosuppression generated by anti-CRISPR proteins. How long will a bacterium remain in this defenceless state? How does this state vary from cell to cell in a bacterial population? Although this cooperation strategy seems to provide one way for phages to neutralize CRISPR–Cas complexes, perhaps other strategies remain to be discovered that do not require the sacrifice of the first wave of infecting phages. ■

Aude Bernheim *and* **Rotem Sorek** *are in the Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel.*
e-mail: rotem.sorek@weizmann.ac.il

1. Borges, A. L. *et al. Cell* https://doi.org/10.1016/j.cell.2018.06.013 (2018).
2. Landsberger, M. *et al. Cell* https://doi.org/10.1016/j.cell.2018.05.058 (2018).
3. Hille, F. *et al. Cell* **172**, 1239–1259 (2018).
4. Borges, A. L., Davidson, A. R. & Bondy-Denomy, J. *Annu. Rev. Virol.* **4**, 37–59 (2017).
5. Pawluk, A., Davidson, A. R. & Maxwell, K. L. *Nature Rev. Microbiol.* **16**, 12–17 (2018).
6. Garneau, J. E. *et al. Nature* **468**, 67–71 (2010).
7. Ptashne, M. *A Genetic Switch* 3rd edn (Cold Spring Harbor Laboratory Press, 2004).
8. Erez, Z. *et al. Nature* **541**, 488–493 (2017).
9. Doron, S. *et al. Science* **359**, eaar4120 (2018).

R.S. declares competing financial interests. See go.nature.com/2l6rtqk for details.

MEDICAL RESEARCH

Histidine degradation boosts cancer therapy

Clinical use of the anticancer drug methotrexate can be limited by its high toxicity. It emerges that a diet rich in the amino acid histidine increases the effectiveness of methotrexate treatment and lowers toxicity in mice. SEE LETTER P.632

CHRISTIAN FREZZA

Methotrexate was one of the first approved anticancer drugs, and is a cornerstone of modern chemotherapy for the treatment of certain solid tumours and blood cancers. However, this therapy must often be stopped prematurely because of its high toxicity in patients' healthy cells. On page 632, Kanarek *et al.*[1] report that degradation of the amino acid histidine can increase the sensitivity of cancer cells to methotrexate, and the results suggest that this approach can also help to lower the side effects of methotrexate treatment in mouse models.

An enzyme cofactor termed folate is required for DNA and protein synthesis. In 1948, it was found[2] that leukaemia is particularly sensitive to drugs that block folate synthesis. Methotrexate inhibits the enzyme dihydrofolate reductase (DHFR), which synthesizes the active form of folate known as tetrahydrofolate (THF) from its precursor dihydrofolate (DHF) (Fig. 1). However, when this drug is used in the clinic, it can substantially reduce the level of folate in healthy cells — a side effect that is only partially relieved by taking folate supplements. If this toxicity reaches a level that compromises the function of healthy cells, therapy may have to be stopped prematurely. There is therefore great interest in trying to reduce the side effects of methotrexate treatment.

To investigate the processes that affect the cellular response to methotrexate, Kanarek *et al.* used a cell line derived from a person with leukaemia. They employed a gene-editing technique that prevents the expression of individual genes throughout the entire genome in such cells, and tested whether a decrease in the expression of any particular gene reduced the cells' sensitivity to methotrexate. One such gene that they identified encodes a protein called SLC19A1, which transports folate

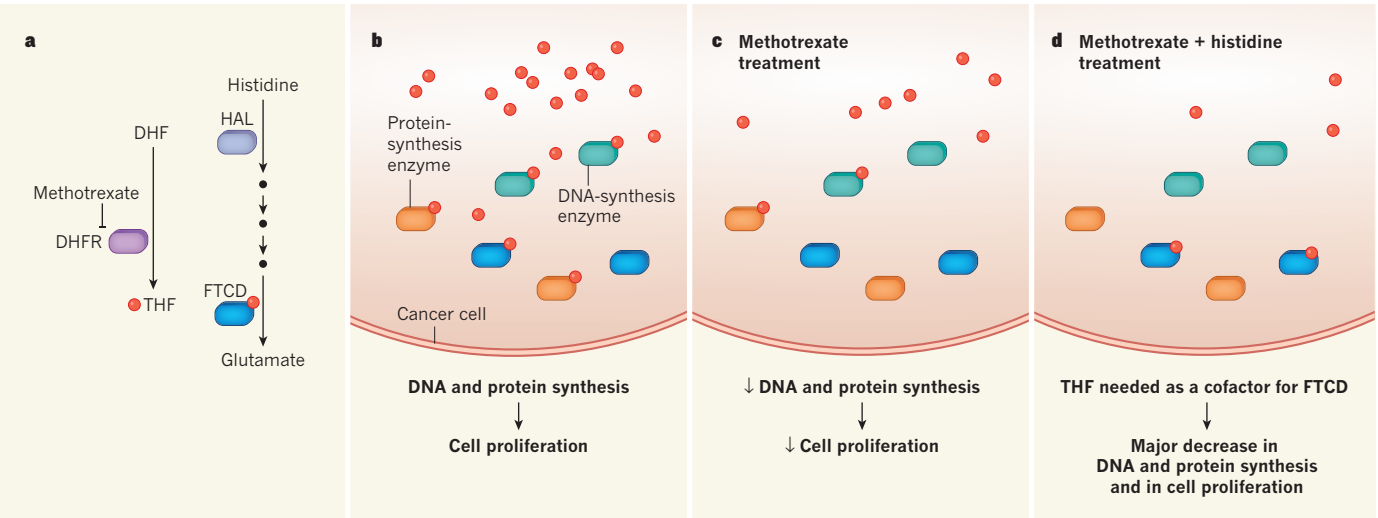


Figure 1 | The effectiveness of the cancer drug methotrexate can be improved by histidine degradation. **a**, The cofactor tetrahydrofolate (THF) is required for the function of enzymes, including those that catalyse processes needed for cell proliferation. The enzyme dihydrofolate reductase (DHFR) generates THF from dihydrofolate (DHF), and this catalytic reaction can be inhibited by the anticancer drug methotrexate. Kanarek *et al.*[1] report that, in mouse models and human cell lines, the sensitivity of cancer cells to methotrexate can be boosted by the metabolic degradation of the amino acid histidine. Histidine is metabolized to the amino acid glutamate through a series of intermediate molecules (black circles) in a pathway that requires the enzyme histidine ammonia lyase (HAL), and, crucially for this effect on methotrexate treatment, the THF-dependent enzyme formimidoyltransferase cyclodeaminase (FTCD). **b**, In cancer cells, a large pool of THF is available for THF-dependent enzymes, such as FTCD or those needed for cell proliferation. **c**, In methotrexate-treated cancer cells, a decrease in THF levels reduces the activity of these enzymes and thereby reduces tumour growth. **d**, The authors report that tumour growth in mice given methotrexate and a histidine-rich diet was substantially decreased compared with **c**. This might be because histidine degradation co-opts THF for use with FTCD, thereby enhancing the effect of methotrexate.

© 2018 Springer Nature Limited. All rights reserved.

and methotrexate into cells. This result was unsurprising, given that SLC19A1 has previously been linked[3] to sensitivity to methotrexate, but it confirmed the power of the authors' approach.

Kanarek and colleagues found that decreasing the expression of a gene that encodes the enzyme formimidoyltransferase cyclodeaminase (FTCD) also reduced the sensitivity of the cancer cells to methotrexate. FTCD degrades histidine, an amino acid that humans must obtain from their diet because the body cannot synthesize it. The authors reasoned that, because FTCD requires THF as a cofactor for its activity, a decrease in the intracellular levels of FTCD could help to limit the reduction of cellular pools of THF that occurs with methotrexate treatment. Consistent with this hypothesis, the authors found that when the levels of FTCD were reduced by gene editing, THF depletion in human cancer cells treated with methotrexate was less than that seen in methotrexate-treated cells in which FTCD levels were not reduced.

On the basis of these findings, the authors predicted that the manipulation of other steps in the histidine-degradation pathway could affect the efficacy of methotrexate treatment. To test this, they again used gene editing to decrease the expression of other genes in this pathway in human cancer cells grown *in vitro*. They found that such decreases did indeed mimic the effect of decreased FTCD levels in preserving the cellular pool of THF in methotrexate-treated cells.

The authors then analysed a panel of cancer cell lines, and observed a correlation between cell lines that had low expression of the enzyme histidine ammonia lyase (HAL), which acts at the first step in the histidine-degradation pathway, and low sensitivity to the effects of methotrexate. Moreover, when Kanarek and colleagues analysed data[4] available for people who have a form of leukaemia that is commonly treated with methotrexate, they found that patients who had high levels of HAL expression had higher survival rates than those with low HAL expression levels. Perhaps this is because methotrexate is more effective when high levels of HAL boost histidine degradation and the depletion of THF pools by FTCD.

The authors therefore proposed that greater cellular histidine degradation should deplete the THF pool and thus increase the effectiveness of methotrexate. To explore this prediction, they generated tumours in mice, and gave the animals methotrexate and a histidine-rich diet. They treated the animals with lower doses of methotrexate than those normally used in such models, to assess whether histidine had a synergistic effect in this context. The histidine-rich diet did indeed boost tumour-growth inhibition in response to low-dose methotrexate compared with the effects of methotrexate in animals that did not receive a dietary histidine boost.

An unexpected and exciting finding of Kanarek and colleagues' work is that, in their mouse models, the toxicity that is usually seen in non-cancerous tissues with methotrexate treatment did not occur when the animals were given a histidine-rich diet, despite the extra toxicity that would have been expected as a result of the depletion of THF in these tissues. The results not only indicate that histidine can increase methotrexate-driven cytotoxicity in cancer cells, but also raise the possibility that some mechanism is reducing the toxic effects of THF depletion in normal tissues. Do normal and tumour tissue differ in sensitivity in this context as a result of having different thresholds for toxicity, or is the rate of histidine degradation different in normal and tumour tissues?

Given the technical challenges involved in measuring certain intermediate molecules in the histidine-degradation pathway, histidine's contribution to the folate pool still needs to be quantified *in vitro* and *in vivo* to validate the authors' model. Furthermore, there are other pathways[5], in addition to that of histidine degradation, for example degradation of the amino acid methionine, that can affect the THF level in cells. It will be important to investigate the extent to which such pathways contribute to the THF pool, and whether they also affect methotrexate-associated toxicity.

Kanarek and colleagues' work indicates the potential value of conducting clinical trials to investigate whether histidine supplementation can boost the effectiveness of methotrexate in cancer treatment. If this is the case, further clinical trials would be needed to determine the optimal dosage. Histidine supplementation might provide an inexpensive and safe strategy for increasing the efficacy of methotrexate. Supplementation with low doses does not have adverse effects in humans, although doses of more than 8 grams of histidine a day can lead to zinc deficiency[6]. However, there have been no extensive studies to investigate possible detrimental effects of a histidine-rich diet in humans.

Finally, these results might have relevance for people with rheumatoid arthritis or other chronic autoimmune diseases, in which low-dose methotrexate is the standard treatment[7] and folate supplements are the only strategy currently available to limit side effects. Could a histidine-rich diet reduce side effects in these people also, and increase their well-being? ■

Christian Frezza *is at the MRC Cancer Unit, University of Cambridge, Hutchison/MRC Research Centre, Cambridge CB2 0XZ, UK.*
e-mail: cf366@mrc-cu.cam.ac.uk

1. Kanarek, N. *et al. Nature* **559**, 632–636 (2018).
2. Farber, S., Diamond, L. K., Mercer, R. D., Sylvester, R. F. Jr & Wolff, J. A. *N. Engl. J. Med.* **238**, 787–793 (1948).
3. Assaraf, Y. G. & Schimke, R. T. *Proc. Natl Acad. Sci. USA* **84**, 7154–7158 (1987).
4. Roberts, K. G. *et al. N. Engl. J. Med.* **371**, 1005–1015 (2014).
5. Locasale, J. W. *Nature Rev. Cancer* **13**, 572–583 (2013).
6. Institute of Medicine of the National Academies. *Dietary Reference Intakes for Energy, Carbohydrate, Fiber, Fat, Fatty Acids, Cholesterol, Protein, and Amino Acids* Ch. 10 (Natl Acad. Press, 2005).
7. Brown, P. M., Pratt, A. G. & Isaacs, R. G. *Nature Rev. Rheumatol.* **12**, 731–742 (2016).

This article was published online on 11 July 2018.

NEUROBIOLOGY

How the head governs the heart of fly species

The males of two closely related species of fly respond differently to a female pheromone. It emerges that this difference is due to alterations in the activity of an evolutionarily conserved neural circuit in the brain. SEE ARTICLE P.564

NICOLAS GOMPEL
& BENJAMIN PRUD'HOMME

Particular sensory signals — a certain odour or taste, for instance — can trigger opposing responses in different species. What changes in neural circuits could elicit such contradictory behaviours? On page 564, Seeholzer *et al.*[1] compare the circuits that control courtship in males in two closely related species of fruit fly that behave differently when exposed to a particular pheromone. They find an explanation for these opposing behaviours that few neuroscientists would have predicted.

While wandering on a rotten apple, a male fruit fly is likely to encounter a variety of other flies, only a few of which will be females of the same species. Identifying and courting these conspecific females is a crucial step in securing the male fly's reproductive success. But this is not always an easy task. For instance, two sister species of fruit fly — *Drosophila melanogaster* and *Drosophila simulans* — are often found on the same fruits. Males of these species therefore rely heavily on sensing pheromones to identify conspecific females[2].

To test whether a female fly is of the same species, the male touches her abdomen with

© 2018 Springer Nature Limited. All rights reserved.

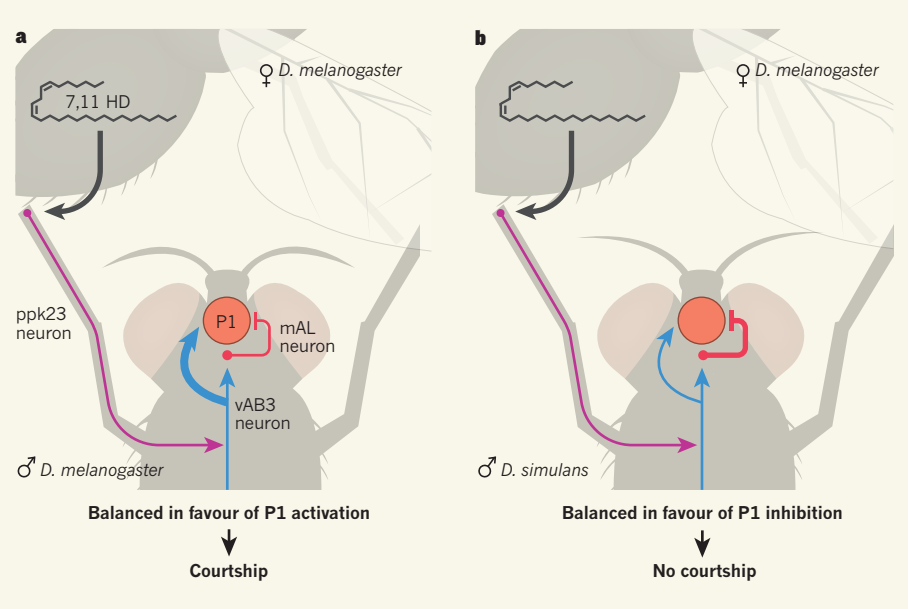


Figure 1 | Changes in the activity of a courtship-promoting neural circuit. **a**, Female fruit flies of the species *Drosophila melanogaster* produce the pheromone 7,11-HD. Seeholzer *et al.*[1] report that *D. melanogaster* males identify females of their species through neurons in the foreleg that express the protein Ppk23. These ppk23 neurons activate a population of neurons called vAB3 that, in turn, activates neurons in the P1 brain region, promoting courtship behaviour. vAB3 neurons also activate the mAL neuron population, which inhibits P1. In this species, vAB3 activation overrides mAL inhibition to trigger P1 activity, leading to courtship. **b**, The authors find that this neural circuit is evolutionarily conserved in males of a sister species, *Drosophila simulans*. However, mAL-mediated inhibition overrides vAB3 activation, so that P1 is inhibited and courtship behaviours are aborted. Whether this difference reflects changes in the activity of mAL neurons or changes in how the inputs are detected by P1 neurons remains unclear.

his foreleg, sampling her pheromones using sensory neurons. *D. melanogaster* females produce the pheromone 7,11-HD, which promotes courtship behaviour when sampled by *D. melanogaster* males[2]. By contrast, *D. simulans* females do not produce 7,11-HD, and *D. simulans* males react to the pheromone (on either *D. melanogaster* females or *D. simulans* females doused with 7,11-HD) by aborting courtship behaviour. Hence, 7,11-HD is detected by males from both species, but elicits contrary behavioural responses in each.

What differences in the nervous systems of the male flies could account for the opposing role of 7,11-HD in regulating courtship decisions in these two species? Seeholzer *et al.* set out to answer this question by comparing the architecture and function of the neuronal circuits that govern courtship behaviour in both. To achieve this, the authors used a combination of sophisticated techniques: the genetic manipulation of specific populations of neurons; the activation of particular neurons with light using a technique called optogenetics; and the observation of neuronal activity in real time using functional imaging.

In *D. melanogaster* males, the decision to initiate courtship is made by the activation of a region in the brain called P1, which contains neurons that integrate multiple sensory signals from outside the brain[3]. Seeholzer *et al.* found that P1 has the same role in *D. simulans*. Moreover, the neural pathway that detects 7,11-HD and propagates the sensory input to P1 is structurally similar in *D. melanogaster* and *D. simulans* (Fig. 1).

The authors showed that 7,11-HD is detected by a few taste-perceiving neurons in the male foreleg of each species that are characterized by expression of the protein Ppk23. These neurons activate a population of neurons called vAB3 that, in turn, directly excites P1 neurons[4]. However, the researchers demonstrated that vAB3 neurons also activate a population of neurons called mAL that directly inhibits P1 neurons[4]. The pheromone-processing pathway therefore bifurcates into excitatory and inhibitory routes that converge on P1. This circuit presumably provides a mechanism for setting pheromone responses and for tightly controlling the excitation of P1.

"The authors have shown clearly that evolutionary changes can be embedded more deeply in brain circuits."

If the structure of the pheromone-processing pathway is evolutionarily conserved between species, how does 7,11-HD elicit a net excitation of P1 to trigger courtship in *D. melanogaster*, but not in *D. simulans*? Seeholzer *et al.* found that it's simply a matter of balance. The mAL-mediated inhibition of P1 neurons completely masks their excitation by vAB3 neurons in *D. simulans*, resulting in a net inhibition of P1. But mAL activity only dampens the activation of P1 by vAB3 in *D. melanogaster*.

Seeholzer and colleagues' work therefore suggests that it is not the architecture of the circuitry that has changed during the evolution of the two species, but the relative weight of the inhibitory and excitatory pathways that converge on P1 neurons. This is unexpected, because the idea that changes outside the brain can explain behavioural diversity has long dominated this field[5–10]. But the authors have shown clearly that evolutionary changes can be embedded more deeply in brain circuits, where no one has looked so far. This is a provocative new conceptual framework that explains how a behaviour can change between closely related species.

The most interesting aspect of Seeholzer and colleague's work is undoubtedly the unexpected nature of the findings. But the study's combination of state-of-the art techniques also makes it a technical triumph, for two reasons. First, using neurogenetic tools designed for the model organism *D. melanogaster* in the non-model species *D. simulans* was a challenge. Second, comparing the structure and function of individual circuit components between species, in the haystack of neurons that makes up the brain, is uncharted territory.

Of course, several questions remain. For instance, what caused the change in balance between the inhibitory and excitatory branches of the courtship pathway? Perhaps it was a progressive change that emerged from genetic variation between individuals belonging to the last common ancestor of these species. Under sexual selection, initial variations that caused two groups of males to respond differently to 7,11-HD, as well as two groups of females to produce different levels of the pheromone, might have gradually built a barrier between the incipient species. The genetic basis of the difference in neural-circuit activity between the two species, however, remains to be defined.

It is also unclear how fine-tuning of the circuit differs between the species. Seeholzer *et al.* showed that vAB3 and mAL neurons are active in both species, so there must be a difference in the way their signalling is propagated to P1. Perhaps there are changes in the levels of signalling by mAL neurons, or maybe P1 neurons respond differently to inhibitory inputs in each species.

Finally, the study highlights that integrative nodes in neural circuits such as P1 might be poised to accommodate evolutionary changes that underlie variations in behaviour. This concept is analogous to the evolution of particular organismal shapes. In such a setting, particular genes at key nodes in gene-regulatory networks are poised to accommodate changes that lead to the generation of forms[11]. Future work, following the approach pioneered by Seeholzer *et al.*, will

© 2018 Springer Nature Limited. All rights reserved.

reveal the extent to which the evolution of behaviour is targeted to particular nodes in neural circuits. ■

Nicolas Gompel *is in the Faculty of Biology, Ludwig Maximilian University, Munich, 82152 Planegg-Martinsried, Germany.* **Benjamin Prud'homme** *is at the Aix-Marseille University, CNRS, Developmental Biology Institute of Marseille (IBDM), 13288 Marseille, France.*
e-mails: gompel@bio.lmu.de; benjamin.prudhomme@univ-amu.fr

1. Seeholzer, L. F., Seppo, M., Stern, D. L. & Ruta, V. *Nature* **559**, 564–569 (2018).
2. Billeter, J. C., Atallah, J., Krupp, J. J., Millar, J. G. & Levine, J. D. *Nature* **461**, 987–991 (2009).
3. Kimura, K., Hachiya, T., Koganezawa, M., Tazawa, T. & Yamamoto, D. *Neuron* **59**, 759–769 (2008).
4. Clowney, E. J., Iguchi, S., Bussell, J. J., Scheer, E. & Ruta, V. *Neuron* **87**, 1036–1049 (2015).
5. Schneiderman, A. M., Hildebrand, J. G., Brennan, M. M. & Tumlinson, J. H. *Nature* **323**, 801–803 (1986).
6. McGrath, P. T. *et al. Nature* **477**, 321–325 (2011).
7. McBride, C. S. *et al. Nature* **515**, 222–227 (2014).
8. Prieto-Godino, L. L. *et al. Nature* **539**, 93–97 (2016).
9. Karageorgi, M. *et al. Curr. Biol.* **27**, 847–853 (2017).
10. Prieto-Godino, L. L. *et al. Neuron* **93**, 661–676 (2017).
11. Stern, D. L. & Orgogozo, V. *Evolution* **62**, 2155–2177 (2008).

This article was published online on 11 July 2018.

PALAEOCLIMATE

Ancient ice sheet had a growth spurt

An analysis of ancient coral from the Great Barrier Reef reveals that global sea level fell rapidly at the end of the last glacial period. The findings suggest that ice sheets are more dynamic than was previously thought. SEE LETTER P.603

PIPPA WHITEHOUSE

The episode known as the Last Glacial Maximum extended from 26,500 to 19,000 years ago, and is often thought of as a prolonged time during which glaciation was continuously at its greatest extent[1]. But on page 603, Yokoyama *et al.*[2] report an analysis of relict coral reefs that suggests that the already extensive ice sheets underwent a final burst of rapid growth about 22,000 years ago, at a rate sufficient to lower global mean sea level by an astounding 17 metres in only 500 years — about ten times faster than the current rate of sea-level rise. This seems to have been followed almost immediately by ice-sheet retreat. Understanding the trigger for this previously undocumented growth is crucial to understanding the sensitivity of ice sheets to external drivers and internal feedbacks.

Previous studies have tended to focus on the magnitude of sea-level decline during the Last Glacial Maximum (LGM), rather than the timing of it, with the general consensus being that sea level was about 125–135 m lower than it is today[3,4]. Geological records that shed light on the detailed timing of sea-level change during this period are notoriously difficult to obtain, not least because they are often located about 100 m underwater. To address this problem, Yokoyama *et al.* analysed a 'staircase' of relict coral reefs that once formed part of the Great Barrier Reef at the edge of the Australian continent — analysis of the recovered corals indicates the sea level at the time they were alive, whereas the reef age was determined using radiometric dating techniques.

One issue with such studies is that modern corals live at a range of depths, creating considerable uncertainty about the depths that can be inferred from fossilized samples[5]. However, by piecing together information from adjacent reefs at different depths, Yokoyama *et al.* built up a composite picture of the timing and magnitude of LGM sea-level decline off the coast of Australia. Their approach is clever and simple: they reason that a hiatus in coral growth caused by sea-level fall at one site must coincide with the development of a second, nearby reef farther offshore, as water depths became shallow enough to support growth at the deeper site (Fig. 1). Moreover, samples recovered from the second site indicate that the sea-level minimum (the lowstand) was short-lived, and that sea level soon began to rise again.

To identify the ice sheets responsible for the dramatic sea-level fall, Yokoyama *et al.* turned to a technique known as glacial isostatic adjustment (GIA) modelling, which can predict how ice-sheet change translates into sea-level change[6]. Although increases in ice mass directly equate to decreases in ocean mass, the resulting pattern of sea-level change is not straightforward. This is because the land beneath ice sheets subsides as mantle material is displaced by the weight of the accumulating ice, and rebounds when the ice melts; an equivalent process takes place across ocean basins in response to increases or decreases in ocean mass. The changes in the shape of the solid Earth alter the shape of the planet's gravity field. And because the gravity field defines the



50 Years Ago

Compensation of £3,200 has so far been paid by the Ministry of Technology for the damage caused by the four sonic boom tests carried out over London last summer with Lightning aircraft. The other seven tests over other parts of Britain cost just over £700 in compensation … With one exception, all the claims involved damage to property. The exception was a claim made by a woman from Hornchurch who suffered a partial loss of hearing … and accepted £150 settlement. Even before the tests, the British Government seems to have recognized that aircraft should not be allowed to fly at supersonic speeds over densely populated areas … It seems inevitable that all supersonic flight over land will be prohibited, at least until a great deal more information on the effect of sonic booms has been collected.
From *Nature* 27 July 1968

100 Years Ago

An article on coal-saving … appears in *Engineering* for July 12. The author … gives average figures for 250 typical steam-boiler plants, covering … 1910 to the present time. It is estimated that 58,500,000 tons of coal per annum are used in this country for steam-raising purposes … exclusive of 15,000,000 tons used in railways. The 250 plants had a total of 1000 boilers, principally of the Lancashire type. With hand-firing the average net working efficiency is 57.8 per cent., as against mechanical firing with an average net working efficiency of 61.4 per cent. … The author estimates that there are 45,000 to 60,000 steam boilers at work in Great Britain, calculated in terms of averaged-sized Lancashire boilers, and considers that all the steam produced in the country to-day could be obtained much more economically with 25 per cent. fewer boilers.
From *Nature* 25 July 1918

© 2018 Springer Nature Limited. All rights reserved.

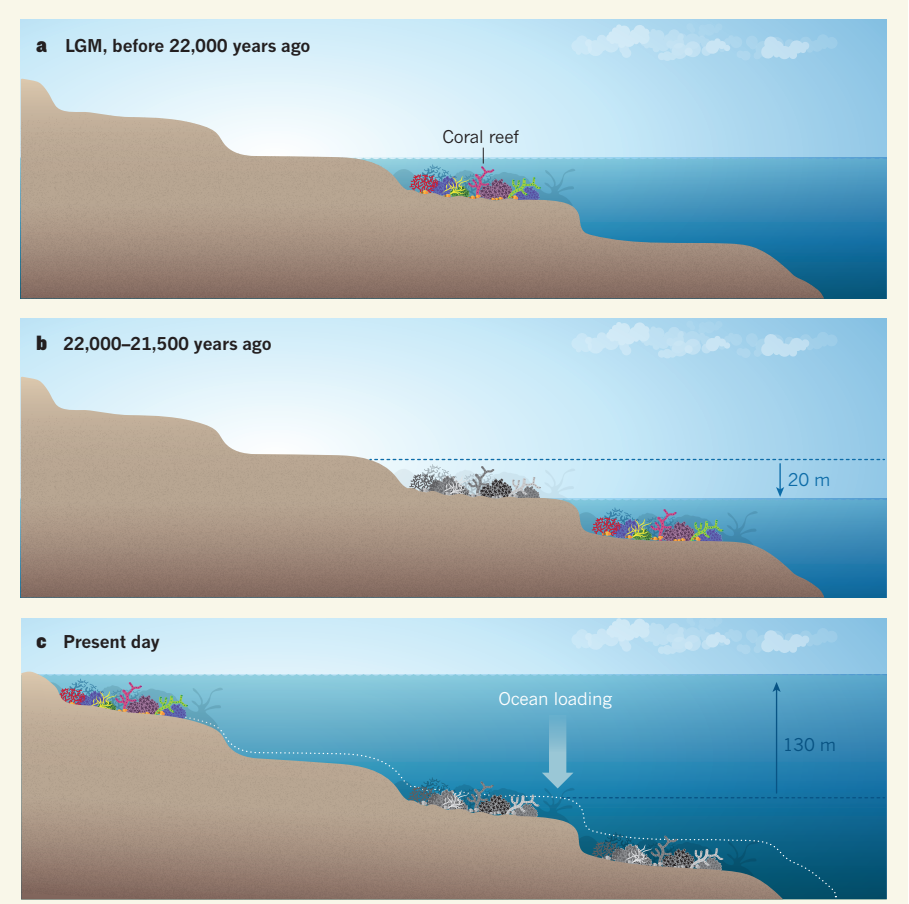


Figure 1 | Rapid changes of sea level recorded by coral reefs. Yokoyama *et al.*[2] studied fossil corals at the Great Barrier Reef to work out how sea level changed during the Last Glacial Maximum (LGM; 26,500 to 19,000 years ago[1]). **a**, Before 22,000 years ago, sea level was about 110 metres below the present level and the reef was confined to a region close to the edge of the continental shelf. **b**, Between 22,000 and 21,500 years ago, a sea-level drop of approximately 20 m altered the position at which coral could grow. The original reef died as it became exposed to air, and a new reef began to grow farther offshore. **c**, The sea surface then rose to present-day levels as ice sheets (not shown) melted. As the ocean flooded the continental shelf, the extra load of water caused the land to tilt (dotted line indicates original position of the continental shelf).

shape of the sea surface, the resulting pattern of sea-level change reflects mass redistribution throughout the Earth system, as well as the net change in ocean mass.

Using a priori assumptions about ice-sheet changes during the last glacial cycle, GIA models can predict the global pattern of sea-level change over time. In places such as Australia, which are far from major ice sheets, sea-level change will be close to the global mean — but even so, a correction must still be applied to account for the tilting of the continental shelf that occurs as it is flooded during post-glacial sea-level rise[7]. By testing different ice-sheet scenarios in a GIA model, and drawing on evidence from other locations that record the sea-level lowstand, Yokoyama *et al.* tentatively propose that growth of the North American ice sheet was the main cause of the rapid sea-level drop at the end of the LGM.

Techniques for reconstructing ice-sheet changes from sea-level observations rely on the fact that sea-level change is not uniform across the planet, but depends on proximity to the changing ice sheets. However, the distribution of currently available sea-level records is insufficient to allow us to determine the precise pattern of ice-sheet change during the LGM, making it difficult to pinpoint the cause of the sudden ice expansion and the reason for its brevity. Additional insight could be gained by studying sites at which sea-level change will be sensitive to changes in nearby ice sheets.

But the interpretation of such 'intermediate field' records would be complex. Most GIA models assume that Earth's properties do not vary from place to place, but analyses of the speed at which seismic waves travel show that this is an over-simplification (see ref. 8, for example). Future sea-level-based inferences of ice-sheet change should use GIA models that incorporate 3D structural variation in their representation of the solid Earth.

Direct evidence of past ice-sheet growth, rather than indirect evidence from ancient shorelines, would also provide crucial insight into the cause of the final LGM sea-level drop reported by Yokoyama and colleagues. Such evidence is, however, extremely difficult to piece together, because growing ice sheets tend to erase evidence of their past extent from the landscape.

Unlike ice-sheet retreat, the mechanisms responsible for rapid ice-sheet growth have received relatively little attention. It is tempting to equate cold periods in Earth's history to an increase in snowfall, but the opposite is actually true, because cold air holds less moisture than warmer air[9]. Yokoyama *et al.* suggest that an increase in the amount of solar radiation received by the Southern Hemisphere might have increased Antarctic snowfall towards the end of the LGM, but increased snowfall is not documented in ice-core records for this period (see ref. 10, for example).

An alternative explanation of Yokoyama and colleagues' findings is that cold ocean temperatures might have led to the expansion and thickening of floating ice adjacent to marine-terminating portions of the major ice sheets. Such buttressing ice shelves have a key role in reducing the flux of ice into the ocean[11], and their presence can promote ice-sheet growth. But reconstructing the extent of ice shelves to investigate this idea will be challenging. Moreover, there is much uncertainty about how ocean temperature changed during the LGM.

Rather than increasing in thickness, did one or more of the ice sheets increase in extent towards the end of the LGM? Lower sea levels and exposed continental shelves would have made it easier for ice sheets to advance offshore at that time, but growing ice sheets depress the land beneath, counteracting the effect of the lower sea level and potentially leading to ice-sheet instability. Could these feedbacks between ice dynamics, the underlying Earth and the adjacent ocean explain the short-lived nature of the final LGM sea-level lowstand, or did external climate factors bring it to an abrupt end? Whatever the explanation, it seems the LGM was more dynamic than previously thought. ■

Pippa Whitehouse *is in the Department of Geography, Durham University, Durham DH1 3LE, UK.*
e-mail: pippa.whitehouse@durham.ac.uk

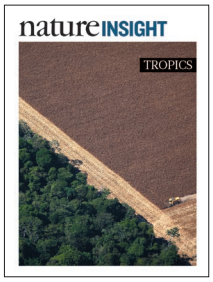
1. Clark, P. U. *et al. Science* **325**, 710–714 (2009).
2. Yokoyama, Y. *et al. Nature* **559**, 603–607 (2018).
3. Peltier, W. R. *Annu. Rev. Earth Planet. Sci.* **32**, 111–149 (2004).
4. Lambeck, K., Rouby, H., Purcell, A., Sun, Y. Y. & Sambridge, M. *Proc. Natl Acad. Sci. USA* **111**, 15296–15303 (2014).
5. Hibbert, F. D. *et al. Quat. Sci. Rev.* **145**, 1–56 (2016).
6. Milne, G. A., Mitrovica, J. X. & Schrag, D. P. *Quat. Sci. Rev.* **21**, 361–376 (2002).
7. Mitrovica, J. X. & Milne, G. A. *Quat. Sci. Rev.* **21**, 2179–2190 (2002).
8. Schaeffer, A. J. & Lebedev, S. *Geophys. J. Int.* **194**, 417–449 (2013).
9. Frieler, K. *et al. Nature Clim. Change* **5**, 348–352 (2015).
10. Fudge, T. J. *et al. Geophys. Res. Lett.* **43**, 3795–3803 (2016).
11. Joughin, I. & Alley, R. B. *Nature Geosci.* **4**, 506–513 (2011).

© 2018 Springer Nature Limited. All rights reserved.

nature INSIGHT

TROPICS

26 July 2018 / Vol 559 / Issue No 7715



Cover image
Rodrigo Baleia

Editor, *Nature*
Magdalena Skipper

Publishing
Richard Hughes

Insights Editor
Ursula Weiss

Subeditors
Rachel Jones, Sarah Farley, Dinah Loon, Mark McGranaghan, David Ribe, Chariklia Rouki, Francisca Schultz, Hugh Williams

Art Editor
Nik Spencer

Sponsorship
Reya Silao

Production
Ian Pope

Marketing
Steven Hurst

Editorial Assistant
Jasmine Delves

The Campus
4 Crinan Street
London N1 9XW, UK
Tel: +44 (0) 20 7833 4000
e: nature@nature.com

SPRINGER NATURE

The tropics are home to a dizzying array of species, store vast quantities of carbon, drive year-to-year climate variability, and will soon be home to half of the world's population. In this Insight, we explore some of the scientific and societal challenges posed by the tropics, and some of the solutions being devised.

Mosquito-borne diseases, such as dengue and malaria, are a major cause of morbidity in the tropics. Neil Ferguson examines how new interventions offer hope of reducing disease burden, but only if supported by rigorous trials and epidemiological analyses.

Ian Gilbert and colleagues explore the challenges associated with developing new drugs to tackle infectious tropical diseases such as malaria and tuberculosis. Recent progress shows that headway can be made with the concerted efforts of all stakeholders.

Non-communicable diseases, such as heart disease and cancer, are often considered a problem of the wealthy. Majid Ezzati and colleagues show that these diseases cause more deaths in low- and middle-income tropical countries than in high-income Western countries.

The tropics harbour more than three-quarters of all species. Jos Barlow and colleagues review the vulnerability of this hyperdiversity to mounting local and global stressors. Only through coordinated local, national and international action can a collapse of tropical biodiversity be prevented, they argue.

In a Review of the tropical forest carbon budget, Edward Mitchard suggests that tropical forests are probably in the process of switching from making an approximately neutral contribution to the global carbon cycle to becoming a net carbon source.

The tropics also set the stage for the El Niño–Southern Oscillation system. Axel Timmermann and colleagues synthesize the complexity of this system, and argue that its variability is excited by a wide range of mechanisms, many of which depend on climatic conditions.

The tropics are undergoing substantial social, environmental and climatic change. Whether an equitable, healthy and sustainable future can be forged, which fosters the region's rich ecological inheritance and the environmental benefits it bestows, will depend on actions taken in the next few years.

Anna Armstrong, I-han Chou, Juliane Mossinger, Clare Thomas, Christina Tobin Kårlström, Michael White
Senior Editors

CONTENTS

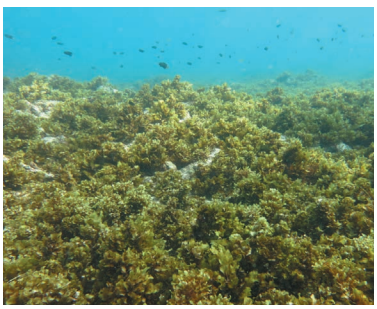
REVIEWS

490 **Challenges and opportunities in controlling mosquito-borne infections**
Neil M Ferguson

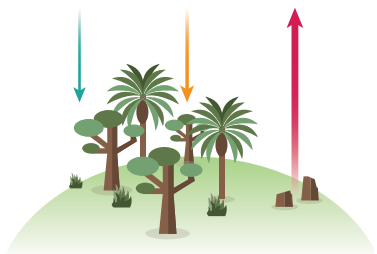
498 **Challenges and recent progress in drug discovery for tropical diseases**
Manu De Rycker, Beatriz Baragaña, Suzanne L Duce & Ian H Gilbert

507 **Acting on non-communicable diseases in low- and middle-income tropical countries**
Majid Ezzati, Jonathan Pearson-Stuttard, James E Bennett & Colin D Mathers

517 **The future of hyperdiverse tropical ecosystems**
Jos Barlow et al.



527 **The tropical forest carbon cycle and climate change**
Edward T A Mitchard



535 **El Niño–Southern Oscillation complexity**
Axel Timmermann et al.

© 2018 Springer Nature Limited. All rights reserved.

REVIEW

https://doi.org/10.1038/s41586-018-0318-5

Challenges and opportunities in controlling mosquito-borne infections

Neil M. Ferguson[1]*

Mosquito-borne diseases remain a major cause of morbidity and mortality across the tropical regions. Despite much progress in the control of malaria, malaria-associated morbidity remains high, whereas arboviruses—most notably dengue—are responsible for a rising burden of disease, even in middle-income countries that have almost completely eliminated malaria. Here I discuss how new interventions offer the promise of considerable future reductions in disease burden. However, I emphasize that intervention programmes need to be underpinned by rigorous trials and quantitative epidemiological analyses. Such analyses suggest that the long-term goal of elimination is more feasible for dengue than for malaria, even if malaria elimination would offer greater overall health benefit to the public.

Deaths from malaria have almost halved since 2000[1–3], despite rapidly growing populations in many endemic regions. Improvements in vector control and (to a lesser extent) treatment are the principal drivers behind this decline[4,5], however, development—leading to improved housing and increased urbanization across the tropics—has also been an important contributing factor[6]. The declines have been particularly marked outside sub-Saharan Africa, with only 10% of malaria deaths now occurring outside this region[7]. Over the same period, the burden of disease from dengue has risen markedly, often in those countries that have the largest reductions in malaria disease[8,9] (Fig. 1). Part of the increase in burden reflects population growth and urbanization in the tropics, but greater connectivity of human populations[10] (which leads to all four dengue serotypes now regularly being detected in all endemic countries[11]), entomology (*Aedes aegypti*, the principal dengue vector, is more highly adapted to urban environments[12]) and climate change (increasing the geographical limits of endemic transmission[13]) have also played a part.

The number of malaria deaths globally remains over 30-fold higher than those from dengue (Fig. 1), although this ratio is less extreme if one only considers disease burden outside Africa. Furthermore, given the availability of highly effective artemisinin-based malaria treatments, a large proportion of remaining mortality induced by malaria reflects gaps in access to treatment or suboptimal diagnosis[4,5]. Conversely, dengue affects urban populations with better (although often not perfect) access to healthcare. No effective dengue antiviral drugs or monoclonal antibodies are currently available for dengue treatment, although improvements in case management have led to substantial reductions in case fatality ratios in situations in which the current best standard of care is available[14]. The unpredictability and explosive nature of dengue epidemics also imposes substantial stresses on healthcare systems, and can cause a lot of anxiety for the public, particularly in contexts in which the burden of disease from most other infections has been markedly reduced in recent decades through vaccination and access to treatment.

Emerging arboviral infections have also caused substantial public concern in recent years[15]. Chikungunya emerged in Latin America and the Caribbean in 2013[16,17], followed by Zika in 2015[18]. Both caused large-scale epidemics over an approximately two-year period, before the accumulation of population immunity in populations across the region led to marked decreases in incidence[19–21]. Despite availability of a highly effective vaccine, yellow fever has also caused relatively large, unpredicted epidemics in Angola and Brazil in the last two years, necessitating rapid large-scale immunization campaigns[22]. However, the true burden of disease caused by both chikungunya and Zika is highly uncertain, in large part because of a lack of systematic surveillance across much of the world and, for Zika, the very mildly symptomatic nature of most infections. The health consequences of infection with both viruses remain poorly characterized and it is unclear to what extent either virus can be considered truly endemic in the human population globally, or whether these viruses, like yellow fever, are dependent on a sylvatic cycle of transmission. These gaps in our knowledge[18] make assessing the public health need for effective interventions highly challenging.

The perceived increasing threat (and disease burden) posed by arboviral infections and the recent global emphasis on reducing the disease burden of malaria has led to increasing investment in the development of new interventions and the intensification of current vector control in many endemic low- and middle-income countries (LMICs). The first vaccines for both malaria and dengue offer some promise for disease reduction but their imperfect and complex efficacy profiles mean that neither represents a panacea and their uptake has therefore been slow[23–27]. However, the accelerated roll-out of long-lived insecticide-treated nets (LLINs) across sub-Saharan Africa over the last decade has led to major declines in malaria incidence, with improved treatment further reducing disease burden[4]. This is not to imply that these gains can be taken for granted—insecticide resistance[28] and/or failure to sustain intervention coverage pose risks of rapid bounce-back. However, by contrast, the evidence that current vector control measures for dengue are having significant impact is limited at best[29–32]. Although differences in the ecology of the principal vectors (for example, in landing periodicity[33]) for the two classes of infections clearly affect the effectiveness of different interventions, I propose that the failure of dengue control to date is principally a consequence of intrinsic differences in the epidemiology of arboviral and malaria infections. These differences necessitate fundamentally different goals for control policy planning for these two classes of infections. I will then review how new technologies for vector control that are currently under development offer the potential to deliver pronounced reductions in disease burden and potentially even elimination in the coming 10–20 years.

Goals of vector control

For all infections, the goals of control are the reduction of disease incidence and, preferably, transmission. Elimination requires

[1]MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London, London, UK. *e-mail: neil.ferguson@imperial.ac.uk

© 2018 Springer Nature Limited. All rights reserved.

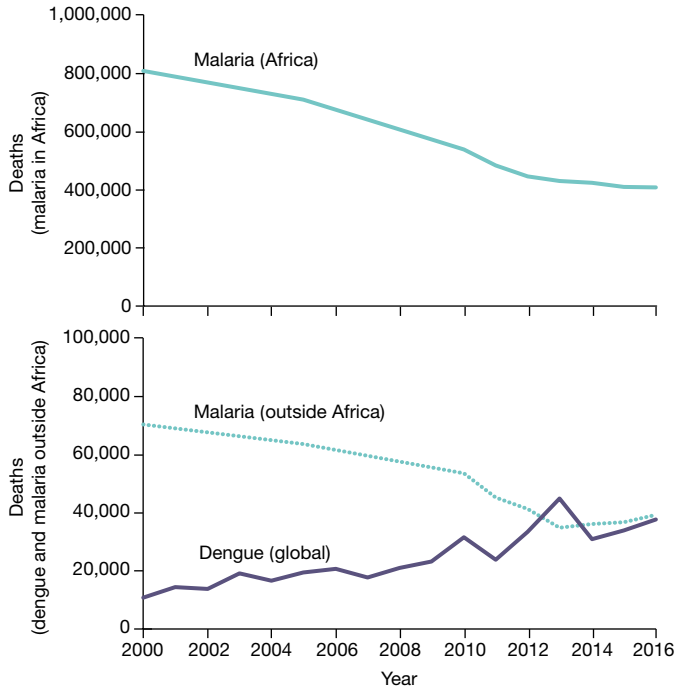


Fig. 1 | Mortality trends between 2000 and 2016 for malaria and dengue. WHO estimates shown for malaria, combining estimates from the 2015 and 2017 World Malaria Reports[3,7]; other estimates[1,8] are substantially higher but show the same overall trends. In the absence of WHO estimates, estimates of dengue-associated mortality from the 2016 Global Burden of Disease study are shown[8]. Top, malaria deaths in Africa. Bottom, dengue deaths and malaria deaths outside Africa.

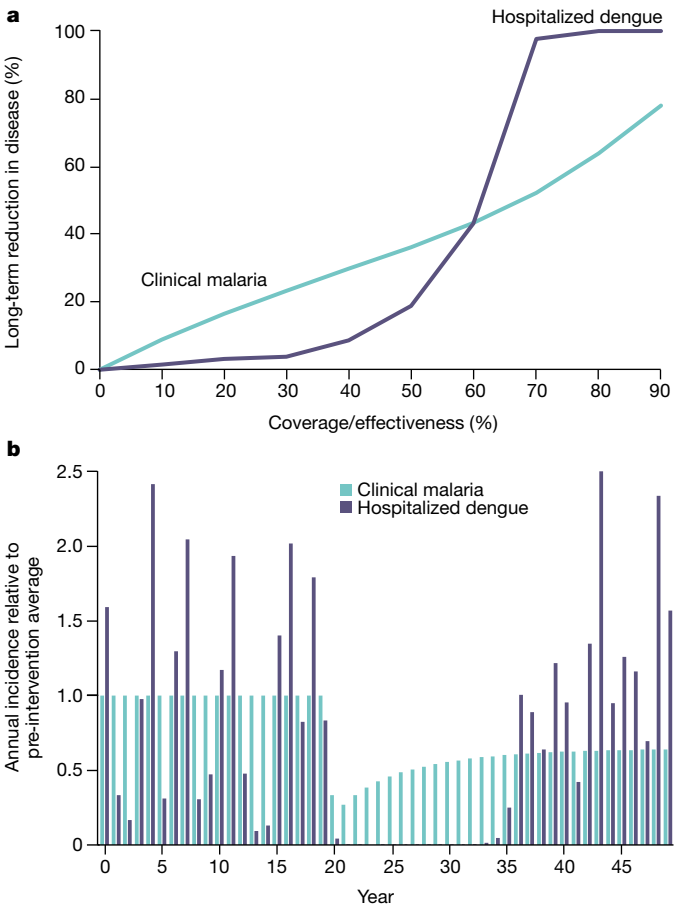


Fig. 2 | Comparative effect of vector controls on incidence of clinical malaria and hospitalized dengue in high transmission intensity settings. a, Published mathematical transmission dynamic models of each disease[23,42] are used to show the long-term effect of varying (i) coverage levels of LLINs for malaria in a sub-Saharan African hyperendemic setting with an (assumed constant) entomological inoculation rate in the absence of controls of 500 (R_0 of approximately 700); (ii) the proportion of dengue exposure blocked by a hypothetical dengue vector control measure (100% coverage assumed) in a setting in which seroprevalence in 9-year olds is 80% on average before the introduction of controls (R_0 of approximately 4). **b**, The temporal impact of controls on annual disease incidence. Controls are introduced in year 20 (50% LLIN coverage assumed for malaria, 50% effective controls for dengue). Inter-annual climate variation is not included in either model, so malaria incidence is constant over time before interventions. Dengue incidence varies markedly year-to-year due to semi-chaotic serotype cycling.

reducing transmission to levels that are less than self-sustaining: in the parlance of mathematical epidemiology, reducing the basic reproduction number (the number of secondary infections caused by a typical index infection in a naive population), R_0, to below 1. Given the lack of vaccines with high efficacy for all mosquito-borne infections other than yellow fever and Japanese encephalitis, the main goal of public health interventions for these infections has been to reduce human exposure through vector control. The rationale for malaria is clear: in high transmission areas, a child might be infected multiple times in a year, with each new infection posing a risk of disease. Reducing exposure by a certain proportion then leads to an immediate and comparable reduction in the incidence of disease—a reduction that can be further improved by improving access to effective antimalarial therapy. Clinical immunity nuances this picture but does not fundamentally change the broad conclusion—although exposures at a later age have a lower risk of disease (severe disease particularly), this risk declines gradually, and both age and exposure have a role in this gradual decline[34].

The same rationale does not hold true for arboviral infections. Unlike malaria (although controversy remains about the extent to which that pathogen can be viewed as a collection of semi-independent antigenically diverse strains[35]), all arboviruses are thought to generate neutralizing homotypic immunity after infection; this means that individuals can experience only one infection with each virus in their lives. Dengue is composed of four distinct (but immunologically cross-reactive) viral serotypes, and therefore four infections are possible. However, secondary dengue infections are responsible for the great majority of severe disease[36–39], with tertiary and quaternary infections thought to be largely asymptomatic[40]. The immunizing nature of arboviral infections has profound consequences for vector control. Completely preventing an individual from being exposed in one year has no impact on lifetime disease risk if a high-level risk of exposure resumes the following year—the only effect of such transient interventions is to postpone infection. The same reasoning applies to partially effective controls. Consider a high transmission setting in which individuals have a 20% risk of dengue infection each year. On average, children will be 5-years old when they experience their first infection, and about 12-years old when they experience their second. Imagine a vector control intervention that reduces exposure by 50%—thus reducing the infection risk to 10% per year. The net long-term effect of this intervention is only to increase the age at which individuals experience their first and second infections—to 10 and 23 years, respectively, for this example. This may paradoxically increase overall disease burden, if dengue disease severity increases with age[41].

This argument can be formalized; Fig. 2a shows the relationship between disease risk and the effectiveness of vector controls at reducing exposure predicted by previously validated mathematical models of dengue and malaria transmission. This malaria model has been used extensively to inform control planning, and was validated against historical prevalence and incidence data both in the absence and presence of control measures[42–44]. The dengue model was previously fitted to the Sanofi–Pasteur dengue vaccine trial data and used to explore the potential impact of large-scale use of the Sanofi–Pasteur vaccine and

© 2018 Springer Nature Limited. All rights reserved.

Wolbachia[23,45]. For malaria, reductions in disease increase almost linearly with coverage. For dengue, the response curve is highly nonlinear, with marginal reductions in lifetime disease risk until the level of exposure reduction is sufficient to reduce R_0 close to 1 (around 70% effectiveness, for this example). I conclude that for vector controls against dengue (or other arboviruses) to have a major long-term impact, they must come close to stopping sustained transmission—that is, achieving elimination. Figure 2a also highlights that elimination, in theory, should require considerably less effort for dengue than for malaria—in high transmission settings, R_0 for dengue is around 4, whereas for malaria it is over 100 (with a large degree of local geographical heterogeneity for both infections[4,46,47]). Therefore, vector control for dengue needs to reduce exposure by 80% to achieve elimination, but for malaria the reduction required is over 99%.

However, focusing only on the long-term effects of interventions neglects the transient temporal effects that the large-scale introduction of an intervention may have on transmission (simulated in Fig. 2b using the models used to generate Fig. 2a). Short- and long-term impacts differ less for malaria than dengue because of the more limited effect of host immunity in modulating disease transmission. The loss of clinical immunity largely explains the rebound in malaria in Fig. 2b. But herd-immunity is fundamental to dengue transmission dynamics; in endemic areas, at any point in time much of the population is immune to any one serotype, so when a new epidemic occurs, it only affects the minority of the population (typically children) who have not yet acquired immunity. Over time, a dynamical equilibrium is reached between viral transmission rates and the level of population immunity, which leads to the effective reproduction number (the average number of secondary infections caused by a typical index case in the presence of population immunity), R, to hover around 1. An exception to this is found in marginally endemic areas that have not yet been affected by all serotypes, where the initial epidemic following invasion with a novel serotype can be much larger than typical[48]; however, following such an initial epidemic, the resulting immunity also causes R for the invading serotype to fall below 1.

In this context, suddenly introducing a new intervention population-wide disrupts this equilibrium, even if the reduction in transmission achieved is insufficient to cause long-term elimination. An intervention that achieves a sustained 20% reduction in exposure would initially reduce R from 1 to 0.8—leading to a temporary cessation of transmission. New births into the population then gradually reduce population immunity, increasing R, and sustained transmission would resume once R once again reached 1. With an annual birth cohort size of, for example, 2% of the total population, it would take up to 10 years for R to increase from 0.8 back to 1. Stopping dengue epidemics for such a time clearly sounds like an impressive outcome, but it is important to note that it would not necessarily lead to a reduction in lifetime disease risk for individuals in the affected population unless more effective interventions could be introduced before the end of the 10-year interim period.

Transient impacts on transmission also need to be taken into account when interpreting data from cluster-randomized trials of vector control interventions. The effect size that is observed in trials of transmission-reduction interventions (such as community mobilization[49]) over timescales of 1 to 2 years will implicitly include the large transient effect on incidence described above, and thus short-term measured effectiveness (for example, reduction in infection rates over one year) would be expected to be considerably greater than the underlying long-term effectiveness of the intervention (as quantified by the fractional reduction in the achieved R_0).

Most of the above arguments around the goals and probable impact of interventions against dengue also hold for chikungunya and Zika, albeit the period between epidemics is much longer and more unpredictable than for dengue[18]. There is one exception, that is, protecting vulnerable populations for a short period. Congenital Zika syndrome is predominantly associated with maternal exposure to Zika in the first trimester of pregnancy[50,51]. Therefore, an intervention that targets individuals (for example, spatial repellents) that reduces exposure for a three-month period might achieve a substantial reduction in the burden of disease caused by a Zika epidemic, even if that intervention has a minimal effect on community levels of transmission. However, just as vector controls that target community transmission require rigorous assessment in randomized trials, so do interventions that aim to reduce personal exposure.

The high level of geographical, environmentally driven heterogeneity in the R_0 of mosquito-borne infections provides nuances, but it does not invalidate these arguments. However, such heterogeneity—and the existence of hotspots with high transmission rates—increases the challenges involved in achieving elimination, albeit with the upside that control in areas with lower transmission rates may be easier than expected from geographically coarse estimates of transmission intensity.

Assessing effectiveness of vector control measures

Why then has there been so little recent success in controlling dengue (especially in LMICs), whereas major reductions in disease for malaria have been achieved? Singapore (which is not a LMIC) is perhaps the only formerly hyper-endemic country in recent decades to have achieved high levels of dengue control (although Cuba achieved similar success in the 1970s)[52]. However, even Singapore has experienced increases in the incidence of dengue incidence in recent years associated with declining levels of population immunity[53]. Nevertheless, substantial reductions in malaria transmission have been achieved in some of the poorest countries of the world in the last decade[4,5].

Part of the answer lies in differences in vector ecology. *Anopheles gambiae*, the primary vector of *Plasmodium falciparum* in sub-Saharan Africa, principally bites at night (and therefore inside houses), which means that LLINs[54,55] (and indoor residual spraying (IRS)[56,57]) are highly effective at reducing human exposure. *A. aegypti* bites during the day (with peaks of activity in the morning and afternoon) both inside and outside houses[33]. Furthermore, the mating patterns and breeding site preferences of *A. aegypti* are better adapted to urban landscapes[58] than most anopheline species (although there is evidence of urban adaptation of the latter[59]).

However, while the ecology of *A. aegypti* limits the effectiveness of 'simple' interventions such as bednets[30], vector control for arboviruses has also suffered from a profound lack of rigorous evidence to support the effectiveness of the individual measures that are currently in use[29–32,52]. In contrast to malaria, for which large cluster-randomized trials with human infection endpoints generated the necessary evidence base to support the scale-up of LLIN and IRS use, randomized trials of vector control approaches for dengue are mostly absent, or have typically been underpowered and only measured entomological rather than epidemiological (that is, human disease) endpoints[32]. 'Integrated vector management' is the recommended approach for vector control of both malaria and dengue, defined as "a rational decision-making process for the optimal use of resources for vector control"[60]. Ironically, the evidence does not currently exist to make rational decisions for dengue vector control, caused by a lack of trial data that support the effectiveness of current control measures at reducing dengue disease.

Most of the current activities to develop vector control strategies against dengue across the tropics are therefore driven more by the understandable hope that reducing mosquito numbers will only benefit public health, than by a quantitative evidence base that directs investments in vector control towards intervention policies that will result in a substantial public health impact. Furthermore, any impact that such interventions might have is often decreased by a tendency of public health authorities to implement vector control measures in response to ongoing dengue epidemics, rather to sustain consistent control all year round. For the reasons outlined in the previous section, reactive short-lived interventions will at best protect individuals transiently, but will leave them susceptible to infection during the next epidemic. Again, Singapore is perhaps the exception—both in terms of the intensity of vector control activities (and the resources made available to sustain

© 2018 Springer Nature Limited. All rights reserved.

these activities) and the mostly consistent, sustained nature of their implementation[53].

In recent years, these shortcomings have been increasingly recognized, culminating in the Vector Control Advisory Group of the WHO (World Health Organization) issuing new guidance that specifies that data from randomized trials with epidemiological endpoints will be required for WHO to recommend new interventions in future[61]. Furthermore, a number of studies have provided improved guidance on appropriate trial design for dengue vector control trials[32,62], highlighting the need for trials to be sufficiently powered to allow for the high degree of spatiotemporal heterogeneity in the incidence of vector-borne diseases, to utilize clusters of sufficient size and spacing to minimize contamination or boundary effects, to have a sufficiently long period of follow-up and to measure epidemiological outcomes. Meeting these requirements is without a doubt challenging, and necessitates that trials are considerably larger (and consequently more expensive) than comparable trials for malaria for which higher and more stable infection rates mean that trials can have a smaller scale.

Indeed, only one dengue vector control intervention study—of the 'Camino Verde' approach to community mobilization[49]—has approached best practice in trial design. However, even in that case, extrapolation of the study results to predict likely long-term, large-scale effectiveness of the intervention is challenging. The small scale of individual clusters (140 households) means that the measured effectiveness may have been affected by individuals living in intervention areas being exposed to infection outside those areas. This 'contamination' issue could indicate that the 25–30% reduction in incidence that is seen in the trial underestimates the true effect size that would be seen if the intervention is implemented at larger scale. Conversely, for the reasons outlined in the previous section, the effect size measured in that trial (and all other short-term trials) over the approximately one-year period of follow-up would be expected to incorporate a short-term perturbative effect of the intervention on dengue transmission, and may therefore overestimate long-term effectiveness.

Assessing the likely long-term impact of interventions from relatively short-term trial data are therefore challenging, not only to adjust for contamination and the typically short-term period of follow-up, but also because the transmission intensity (R_0) of all mosquito-borne infections show high levels of spatiotemporal variation[4,46,63]. Therefore, the effect size seen in one context may overestimate the effectiveness of the same intervention applied in a higher transmission intensity context, or underestimate the effects of the intervention in a lower transmission intensity setting. Repeating studies in a range of contexts can mitigate this issue, but this is costly. Mathematical modelling of each trial incorporating details of the transmission context, period of follow-up and cluster size is therefore required to derive estimates of effectiveness that can be used to predict the likely long-term impact of large-scale use of novel interventions across a variety of transmission contexts. Such analyses are not straightforward (and are not a substitute for long-term follow-up data on effectiveness), but have been undertaken for LLINs and IRS for malaria[42,44], and for both the CYD-TDV (also known by the trade name Dengvaxia, developed by Sanofi–Pasteur) dengue[23,24] and RTS,S (also known by the trade name Mosquirix, developed by GlaxoSmithKline)[26,64,65] malaria vaccines, and have been planned for the analysis of trials of *Wolbachia* as a dengue control measure[45,62].

Intervention trials for Zika and chikungunya—whether of vector controls, vaccines or treatment—are even more challenging than for dengue[66], because of the longer inter-epidemic period and the highly unpredictable nature of epidemic timing. The traditional trial design of recruiting cohorts of participants, randomizing to intervention or control and monitoring outcomes is therefore likely to be prohibitively expensive—owing to the very large numbers of participants and sites needing to be included. Although model-based analysis of available surveillance data can improve the efficiency of site selection[67], reactive designs may be more efficient long-term[68]. Such a trial might involve gaining ethical and regulatory approval to proceed in advance for a large number of potential trial sites, but only triggering recruitment of participants and intervention implementation at a site once transmission is detected there. That said, given their shared vectors and similar transmissibility, vector control interventions that show high efficacy against dengue are likely to also show comparable efficacy against Zika and chikungunya.

The promise of new interventions

Although the large-scale roll-out of LLINs (together with targeted use of IRS) across sub-Saharan Africa in the last decade has had a major effect on both malaria transmission and disease, even very high coverage levels of these interventions and effective treatment are predicted to be insufficient to eliminate malaria from the highest transmission settings[42,44]. Insecticide resistance also poses a growing threat[28]. Mass drug administration can have a very large short-term effect on malaria prevalence[69] (by reducing the parasite reservoir in humans), but unless administration at high coverage levels is repeated indefinitely, such effects are transient, since transmission quickly restarts once infection is reimported from untreated areas[70]. Similarly, seasonal malaria chemoprevention can be highly effective at reducing disease incidence[71] but is only appropriate for moderate to high transmission settings with highly seasonal transmission. Furthermore, both mass drug administration and seasonal malaria chemoprevention pose the risk of accelerating the development of wide-scale artemisinin drug resistance in the parasite population.

For dengue, the situation is worse. As described above, there are currently no interventions in large-scale use in LMICs that are likely to be having any major impact on dengue transmission or disease burden[52]. The intensive vector control interventions adopted largely successfully by Singapore (and high-income settings such as Florida and Queensland) are not easily translated to most LMICs and are likely to be unaffordable for those countries.

In addition to further evaluation and scale-up of current interventions, new intervention technologies are therefore needed to make elimination a feasible policy goal for either dengue or malaria. Over the last two decades, major investment and effort has been committed to the development of both vaccines and novel vector controls, with varying degrees of success.

Vaccines

The CYD-TDV dengue and RTS,S malaria vaccines, the first to be licensed for either infection, are both the result of over twenty years of development effort. Past development efforts have been hindered by the antigenic/immunological complexity of the pathogens and the lack of an obvious commercial market that is sufficient to justify development costs.

In the phase III trial, over two years of follow-up, RTS,S showed 28% (95% confidence interval, 22–33%) efficacy against clinical malaria in children who received three doses over a three-month period, and 36% (95% confidence interval, 31–41%) efficacy in children who additionally received a month-20 booster dose[27]. Efficacy was a little lower in infants and against severe malaria. Data from phase II and III studies and associated mathematical modelling of these data demonstrated waning of efficacy over time[25,27], which was correlated with a decrease in anti-circumsporozoite antibody titres[65]. This modelling also showed that anti-circumsporozoite antibody titres were a correlate of vaccine-induced protection[64,65]. The model of vaccine action proposed by these analyses also explained the observed negative correlation between vaccine efficacy and the transmission intensity seen at trial sites, and the faster decay of efficacy seen in higher transmission settings.

For CYD-TDV, two large phase III studies (in Asia and Latin America) both showed approximately 60% efficacy against virologically confirmed clinical dengue disease in the one year following completion of a three-dose vaccine schedule[72,73]. Efficacy varied by serotype, increased with age and was higher for severe dengue. In the

© 2018 Springer Nature Limited. All rights reserved.

immunological subset of trial participants for whom dengue serological status was measured before the first dose, vaccine efficacy was approximately 75% in individuals who were dengue seropositive, but much lower (and non-significant) in individuals who were dengue seronegative. However, in the first year of long-term follow-up, a statistically significant excess risk of hospitalized dengue disease was observed in vaccine recipients in the youngest age group (2–5 years old). Since no evidence of excess risk was seen in children over 9 years of age in either trial, Sanofi–Pasteur proceeded with submitting the vaccine for regulatory approval with an age indication of use only in children over 9. The most parsimonious and plausible hypothesis that explains these results is that the vaccine acts akin to a silent natural dengue infection[23]. Seronegative recipients of the vaccine are immunologically primed (akin to natural primary infection), so that their first breakthrough natural infection has the higher severity associated with natural secondary infection in unvaccinated individuals. Conversely, seropositive recipients of the vaccine will have their antibody titres against all four serotypes boosted to the high levels seen after secondary infection in unvaccinated individuals—leading to the first breakthrough infection having the low severity associated with natural tertiary infection.

Mathematical modelling was key to extrapolating from the trial results to assess the likely public health impact of large-scale use of both vaccines, and these analyses had a key role in informing WHO recommendations. For RTS,S, modelling predicted that the vaccine could prevent 1 malaria death per 200 vaccine recipients in moderate to high endemicity settings, making it highly cost-effective compared with many other vaccines[26], albeit less cost-effective than LLINs[74]. WHO recognized the notable potential public health benefits offered by the vaccine, but because of the potential difficulty of delivering a four-dose vaccine schedule and in light of the meningitis safety signal, recommended that large-scale pilot implementation programmes should be conducted[75], which are now due to start this year in Ghana, Kenya and Malawi.

For CYD-TDV, modelling indicated that large-scale vaccination might reduce the incidence of symptomatic and hospitalized dengue disease by up to 25% in high transmission intensity settings. However, vaccination was predicted to potentially increase the incidence of hospitalized dengue in low transmission intensity (and thus low seroprevalence) settings[23,24]. WHO recommendations reflected this risk, suggesting that seroprevalence surveys across the population should be undertaken to assess transmission intensity before vaccination roll-out, and that the vaccine should only be used in settings in which over 70% of vaccine recipients were likely to be seropositive[76]. These recommendations have now changed to recommend vaccination only in individuals who test seropositive in light of recent data collected by Sanofi–Pasteur that has conclusively demonstrated that seronegative recipients of all ages experience a higher risk of hospitalized dengue disease throughout the long-term follow-up of the trials[77].

A number of next-generation dengue and malaria vaccines are currently under development. Two other tetravalent live-attenuated dengue vaccines are currently in phase III trials, with initial results due in the next 12 months. It is unclear whether either will also pose risks of use in seronegative recipients that are similar to CYD-TDV, but similarities in the immunogenicity profiles of all three vaccines makes this a possibility[78]. A variety of next-generation pre-erythrocytic, blood-stage and transmission-blocking malaria vaccines are under clinical development[79], but none have yet entered phase III studies. For both infections, it is therefore arguably unlikely that vaccines will become available in the next decade that can on their own offer the promise of disease elimination. However, that is not to say that vaccines will not have an impact in that time frame: both current vaccines and next generation late-stage candidates may make a considerable contribution to reducing disease burden, combinations of vaccines targeting different parasite life stages may offer synergistic levels of protection against malaria, and some potential exists for the development of 'universal' dengue vaccines[80].

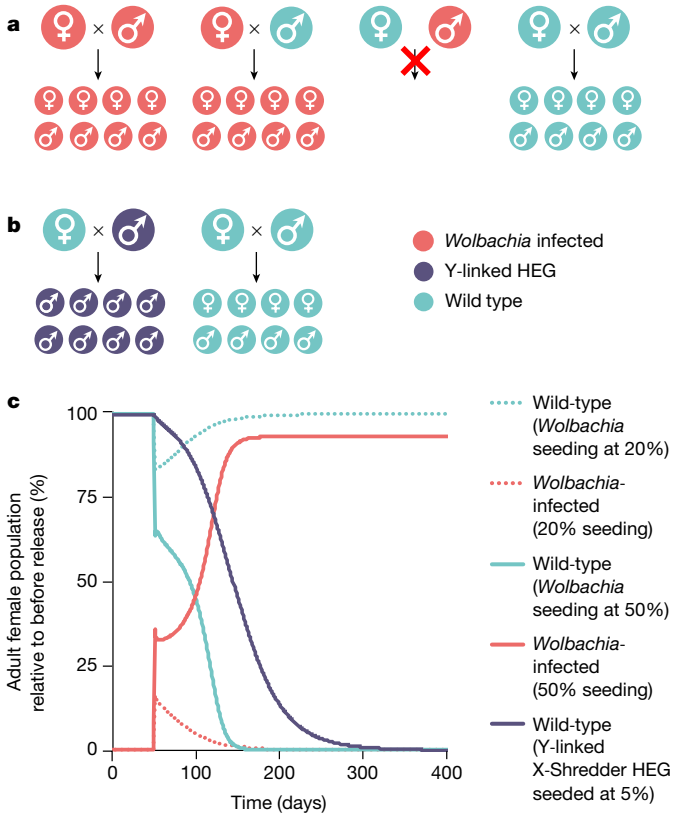


Fig. 3 | Mechanism of action of *Wolbachia* and the Y-linked 'X-shredder' homing-endonuclease gene-based gene drive control measures. **a**, *Wolbachia*-infected mosquitoes are refractory to arboviruses and are able to invade wild-type populations owing to cytoplasmic incompatibility, which leads to crosses between wild-type females and *Wolbachia*-infected males that are non-viable. **b**, A Y-linked X-shredder homing-endonuclease gene (HEG) that induces male mosquito survival biases the sex ratio, since a high proportion (here shown as all) of progeny in their crosses with wild-type females are male and will inherit the homing endonuclease genes. **c**, Illustrative invasion dynamics (releases occurring on day 50). *Wolbachia* exhibits frequency dependent invasion dynamics, with a threshold frequency determined by fitness costs of infection (here assumed to be 30%). Therefore, invasion can occur when the release size equals 50% of the resident wild-type population, but not at 20%. Mosquitoes that carry Y-linked X-shredder homing endonuclease genes can invade at any frequency (here assumed to be 5% of resident population size) and cause the adult mosquito population to crash.

Novel vector control technologies

New insecticides and delivery systems continue to be developed; and perhaps the most interesting and potentially transformative are those interventions that target obligate life-cycle stages such as sugar or blood feeding. Although not a new idea, attractive toxic sugar bait technology[81]—which targets sugar-feeding and therefore potentially increases mortality in both male and female mosquitoes—has been advancing rapidly in recent years, with very promising results (at least in anopheline species) seen in recent small-scale trials[82]. Similarly, ivermectin (and newer longer-lived mosquitocidal drug candidates) could cause substantial suppression of mosquito populations and thus malaria transmission if used as part of a mass treatment intervention[83]. However, the approaches to vector control that offer the potential of transformative impacts are those that may give long-term (or even permanent) reductions in disease transmission after only a single implementation period, either by reducing mosquito density or reducing vector competence. Two such technologies are under active development: *Wolbachia* and gene-drive approaches to genetic modification of mosquito species (Fig. 3).

Wolbachia is a genus of bacteria that naturally infects many insects[84], strains of which have been deliberately transfected into *A. aegypti*

© 2018 Springer Nature Limited. All rights reserved.

mosquitoes. *Wolbachia* typically transmits vertically by achieving high densities in insect eggs. It confers a frequency-dependent fitness advantage relative to uninfected insects through a mechanism of cytoplasmic incompatibility, which results in crosses between *Wolbachia*-infected male insects and wild-type females (which would otherwise result in uninfected progeny) that are unviable (Fig. 3a). As a biological vector control measure, this offers the advantage that following initial releases of *Wolbachia*-infected mosquitoes into a wild-type population, the frequency of *Wolbachia* infection will rise to very high levels as the released mosquitoes interbreed with wild-type insects (Fig. 3c). Initially, *Wolbachia* was envisaged as a means to reduce mosquito density, by using a strain (*w*MelPop) that imposed life-shortening fitness costs on mosquito hosts. However, *Wolbachia*-infected *A. aegypti* were also found to have substantially lower vector competence for a broad range of arboviruses[85], including dengue[86]. It is this phenotype that is being exploited—using the less pathogenic *w*Mel strain—by the World Mosquito Program (https://www.worldmosquitoprogram.org/), the leading development project for *Wolbachia* technology in *A. aegypti*. Mathematical modelling of data from experimental dengue-infection studies in *w*Mel-infected and wild-type *A. aegypti* suggests that successful large-scale release and establishment of *w*Mel-infected mosquitoes could reduce dengue transmission intensity (R_0) by 75%[87]—sufficient to achieve elimination of dengue transmission for decades in even high transmission settings, and permanent elimination in low to moderate transmission settings[45]. Following an extensive programme of small-scale releases, a cluster randomized trial with epidemiological endpoints is now underway in Yogyakarta, Indonesia, with larger non-randomized pilot release studies ongoing in Medellin, Colombia and Rio de Janeiro, Brazil (World Mosquito Program (https://www.worldmosquitoprogram.org/)). Although at a much earlier stage, Wolbachia may also have some potential as a malaria control measure[88].

Gene-drive systems for genetically modified mosquitoes[89,90] offer similar advantages to *Wolbachia* in potentially allowing finite releases of modified mosquitoes to invade wild-type populations and rise to high frequencies. A variety of gene drive systems have been explored experimentally and theoretically[91,92], but the two most developed technologies are homing endonuclease genes and CRISPR–Cas9. Both rely on homology-dependent repair to be copied from one chromosome to its homologous chromosome during mitosis of germline cells or meiosis[89,90]. In the vector control context, most attention to date has been given to engineering constructs that supress mosquito populations. Biasing the sex ratio towards males is one approach to achieving this. So-called X-shredder constructs achieve this by expressing endonucleases that are only expressed during spermatogenesis and that cleave the X chromosome, rendering it non-functional[93]. Inserting such constructs into the Y chromosome would generate a highly invasive genetically modified mosquito that in theory could be introduced into a wild-type population once and would eventually drive that population to extinction[94,95] (Fig. 3b, c). In *Aedes aegpti*, targeting male-determining factors using germline Cas9 expression could give similar results[91,96], although research is at an earlier stage compared with anopheline systems. Development of constructs that reduce vector competence is another area of active research[89], and a variety of potential targets have been identified for both malaria[97,98] and arboviruses[99].

Both *Wolbachia* and gene-drive technologies face several challenges. First, gaining public acceptance and regulatory approval requires a rigorous risk assessment/management process and intensive stakeholder engagement. Such barriers are clearly higher for genetically modified organisms than purely biological control measures such as *Wolbachia*. Second, use of these interventions will impose intense selection pressures that are likely to drive the evolution of resistance in either the target pathogen (in the case of *Wolbachia* or vector competence gene drive constructs) or the vector (for population suppression genetic constructs, and perhaps for *Wolbachia*), and the long-term phenotypic stability of *Wolbachia* in *A. aegypti* is yet to be determined. Although several modelling studies have been undertaken, knowledge gaps in vector ecology (for example, regarding the intensity of density-dependent regulation of larval populations, the extent and nature of overwintering mechanisms) make prediction of the impact of these interventions uncertain, especially for gene-drive systems.

Towards elimination

Although current interventions have proven insufficient so far, the novel vector control technologies described above may make elimination of dengue or malaria a feasible goal, even in the highest transmission settings. However, despite the greater progress in controlling malaria than dengue in recent decades, the scale of the challenge to eliminate malaria from the highest transmission hyperendemic settings is larger. The R_0 of dengue probably rarely exceeds 6, which means that reducing transmission by approximately 85% should be sufficient for permanent elimination in nearly all settings. Plus, for the reasons outlined above, even interventions that achieve a (sustained) 30% reduction in R_0 should cause marked reductions in dengue incidence for decades if applied at sufficient scale—buying time for more effective interventions to be developed. By contrast, the R_0 for *P. falciparum* is in the hundreds in the highest transmission settings, requiring a >99% reduction in transmission intensity to achieve elimination. In addition, there are multiple *Plasmodium* species that are capable of causing disease in humans, and multiple competent vector species even outside the *A. gambiae* complex. Nevertheless, even if malaria elimination (and eventual eradication) remains a more distant goal, the vector tools that are currently under development offer the promise of delivering order-of-magnitude reductions in transmission and disease.

While elimination may be the long-term goal of control programmes, disease reduction must remain the priority in the short to medium term. In this context, a wider range of current and near-to-market interventions are relevant—both vector control measures and vaccines. One of the greatest challenges will be to make optimal use of limited resources (especially in low-income settings) to deliver the greatest public health impact. Rigorous epidemiological analysis and mathematical modelling will be key to ensuring such optimal deployment—for the extrapolation of clinical trial data to predict population impact of each intervention in a wider range of settings and in combination with other control measures. Rigorous monitoring and evaluation is also essential to evaluate the real-world effectiveness of interventions. The epidemiology, vector ecology and transmission intensity of both arboviruses and malaria is highly geographically variable, requiring intervention policies to be tailored to the local environment. Modelling has been pivotal in facilitating the characterization of this heterogeneity[4,46].

There is unlikely to be a single 'silver bullet' intervention that is sufficient to control either class of infections, so policy formulation will require setting-specific selection of interventions from the overall portfolio of available tools that have good evidence supporting their effectiveness. Such optimization should account for the potential synergistic benefits of combining interventions with different mechanisms of action (for example, vector control and vaccines). Intervention effectiveness is only one of the criteria relevant to making that selection—cost, ease of delivery and public acceptability (and thus achievable coverage) are equally important. Such optimization is now possible using integrated disease transmission and health economic models[44,74], although reliable surveillance and intervention effectiveness data are essential to calibrate such models. Last, data analytics (spanning dynamical modelling, statistical analysis and machine learning) will become increasingly critical to cope with the ever-growing volumes of surveillance, genomic, remote sensing and other (for example, mobile phone[100]) data becoming available—to synthesize multiple data streams and derive actionable insights to inform public health policy-making.

Received: 19 February 2018; Accepted: 24 May 2018;
Published online 25 July 2018.

1. Gething, P. W. et al. Mapping *Plasmodium falciparum* mortality in Africa between 1990 and 2015. *N. Engl. J. Med.* **375**, 2435–2445 (2016).
2. Murray, C. J. L. et al. Global malaria mortality between 1980 and 2010: a systematic analysis. *Lancet* **379**, 413–431 (2012).

© 2018 Springer Nature Limited. All rights reserved.

3. WHO. Estimated malaria deaths by region. http://www.who.int/gho/malaria/epidemic/deaths/en/ (2016).
4. Bhatt, S. et al. The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. *Nature* **526**, 207–211 (2015).
This study estimates the impact of different malaria interventions and treatment on infection prevalence and disease incidence.
5. Cibulskis, R. E. et al. Malaria: global progress 2000–2015 and future challenges. *Infect. Dis. Poverty* **5**, 61 (2016).
6. Tatem, A. J., Gething, P. W., Smith, D. L. & Hay, S. I. Urbanization and the global malaria recession. *Malar. J.* **12**, 133 (2013).
7. WHO. World malaria report 2017. http://www.who.int/malaria/publications/world-malaria-report-2017/en/ (2017).
8. GBD 2016 Causes of Death Collaborators. Global, regional, and national age–sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet* **390**, 1151–1210 (2017).
9. Stanaway, J. D. et al. The global burden of dengue: an analysis from the Global Burden of Disease Study 2013. *Lancet Infect. Dis.* **16**, 712–723 (2016).
10. Tian, H. et al. Increasing airline travel may facilitate co-circulation of multiple dengue virus serotypes in Asia. *PLoS Negl. Trop. Dis.* **11**, e0005694 (2017).
11. Messina, J. P. et al. Global spread of dengue virus types: mapping the 70 year history. *Trends Microbiol.* **22**, 138–146 (2014).
12. Jansen, C. C. & Beebe, N. W. The dengue vector *Aedes aegypti*: what comes next. *Microbes Infect.* **12**, 272–279 (2010).
13. Campbell, L. P. et al. Climate change influences on global distributions of dengue and chikungunya virus vectors. *Phil. Trans. R. Soc. B* **370**, 20140135 (2015).
14. Kalayanarooj, S., Rothman, A. L. & Srikiatkhachorn, A. Case management of dengue: lessons learned. *J. Infect. Dis.* **215**, S79–S88 (2017).
15. Weaver, S. C., Charlier, C., Vasilakis, N. & Lecuit, M. Zika, chikungunya, and other emerging vector-borne viral diseases. *Annu. Rev. Med.* **69**, 395–408 (2018).
16. Leparc-Goffart, I., Nougairede, A., Cassadou, S., Prat, C. & de Lamballerie, X. Chikungunya in the Americas. *Lancet* **383**, 514 (2014).
17. Pialoux, G., Gaüzère, B.-A., Jauréguiberry, S. & Strobel, M. Chikungunya, an epidemic arbovirosis. *Lancet Infect. Dis.* **7**, 319–327 (2007).
18. Lessler, J. et al. Assessing the global threat from Zika virus. *Science* **353**, aaf8160 (2016).
A comprehensive review of the epidemiology of Zika, the public health threat posed by Zika infections and current uncertainties.
19. Cohen, J. Where has all the Zika gone? *Science* **357**, 631–632 (2017).
20. Ferguson, N. M. et al. Countering Zika in Latin America. *Science* **353**, 353–354 (2016).
21. Netto, E. M. et al. High Zika virus seroprevalence in Salvador, Northeastern Brazil limits the potential for further outbreaks. *MBio* **8**, e01390-17 (2017).
22. Shearer, F. M. et al. Existing and potential infection risk zones of yellow fever worldwide: a modelling analysis. *Lancet Glob. Health* **6**, e270–e278 (2018).
23. Ferguson, N. M. et al. Benefits and risks of the Sanofi–Pasteur dengue vaccine: modeling optimal deployment. *Science* **353**, 1033–1036 (2016).
24. Flasche, S. et al. The long-term safety, public health impact, and cost-effectiveness of routine vaccination with a recombinant, live-attenuated dengue vaccine (Dengvaxia): a model comparison study. *PLoS Med.* **13**, e1002181 (2016).
This study presents the model comparison exercise that informed initial WHO recommendations on the use of the Sanofi dengue vaccine.
25. Olotu, A. et al. Seven-year efficacy of RTS,S/AS01 malaria vaccine among young African children. *N. Engl. J. Med.* **374**, 2519–2529 (2016).
26. Penny, M. A. et al. Public health impact and cost-effectiveness of the RTS,S/AS01 malaria vaccine: a systematic comparison of predictions from four mathematical models. *Lancet* **387**, 367–375 (2016).
This study presents the model comparison exercise that informed WHO recommendations on the use of the RTS, S malaria vaccine.
27. RTS,S Clinical Trials Partnership. Efficacy and safety of RTS,S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: final results of a phase 3, individually randomised, controlled trial. *Lancet* **386**, 31–45 (2015).
28. Hemingway, J. et al. Averting a malaria disaster: will insecticide resistance derail malaria control? *Lancet* **387**, 1785–1788 (2016).
29. Achee, N. L. et al. A critical assessment of vector control for dengue prevention. *PLoS Negl. Trop. Dis.* **9**, e0003655 (2015).
30. Bowman, L. R., Donegan, S. & McCall, P. J. Is dengue vector control deficient in effectiveness or evidence?: Systematic review and meta-analysis. *PLoS Negl. Trop. Dis.* **10**, e0004551 (2016).
31. Heintze, C., Velasco Garrido, M. & Kroeger, A. What do community-based dengue control programmes achieve? A systematic review of published evaluations. *Trans. R. Soc. Trop. Med. Hyg.* **101**, 317–325 (2007).
32. Wilson, A. L. et al. Evidence-based vector control? Improving the quality of vector control trials. *Trends Parasitol.* **31**, 380–390 (2015). This paper outlines good practice for the design of field trials of vector control interventions.
33. Barrozo, R. B., Schilman, P. E., Minoli, S. A. & Lazzari, C. R. Daily rhythms in disease-vector insects. *Biol. Rhythm Res.* **35**, 79–92 (2004).
34. Griffin, J. T. et al. Gradual acquisition of immunity to severe malaria with increasing exposure. *Proc. R. Soc. B* **282**, 20142657 (2015).
35. Gupta, S., Trenholme, K., Anderson, R. M. & Day, K. P. Antigenic diversity and the transmission dynamics of *Plasmodium falciparum*. *Science* **263**, 961–963 (1994).
36. Clapham, H. E., Cummings, D. A. T. & Johansson, M. A. Immune status alters the probability of apparent illness due to dengue virus infection: evidence from a pooled analysis across multiple cohort and cluster studies. *PLoS Negl. Trop. Dis.* **11**, e0005926 (2017).
37. Guzman, M. G., Alvarez, M. & Halstead, S. B. Secondary infection as a risk factor for dengue hemorrhagic fever/dengue shock syndrome: an historical perspective and role of antibody-dependent enhancement of infection. *Arch. Virol.* **158**, 1445–1459 (2013).
38. Katzelnick, L. C. et al. Antibody-dependent enhancement of severe dengue disease in humans. *Science* **358**, 929–932 (2017).
39. Salje, H. et al. Reconstruction of antibody dynamics and infection histories to evaluate dengue risk. *Nature* **557**, 719–723 (2018).
40. Gibbons, R. V. et al. Analysis of repeat hospital admissions for dengue to estimate the frequency of third or fourth dengue infections resulting in admissions and dengue hemorrhagic fever, and serotype sequences. *Am. J. Trop. Med. Hyg.* **77**, 910–913 (2007).
41. Thai, K. T. D. et al. Age-specificity of clinical dengue during primary and secondary infections. *PLoS Negl. Trop. Dis.* **5**, e1180 (2011).
42. Griffin, J. T. et al. Potential for reduction of burden and local elimination of malaria by reducing *Plasmodium falciparum* malaria transmission: a mathematical modelling study. *Lancet Infect. Dis.* **16**, 465–472 (2016).
43. Griffin, J. T. et al. Reducing *Plasmodium falciparum* malaria transmission in Africa: a model-based evaluation of intervention strategies. *PLoS Med.* **7**, e1000324 (2010).
44. Walker, P. G. T., Griffin, J. T., Ferguson, N. M. & Ghani, A. C. Estimating the most efficient allocation of interventions to achieve reductions in *Plasmodium falciparum* malaria burden and transmission in Africa: a modelling study. *Lancet Glob. Health* **4**, e474–e484 (2016).
This study uses mathematical modelling to assess optimal allocation of different malaria interventions in different transmission settings across Africa.
45. Dorigatti, I., McCormack, C., Nedjati-Gilani, G. & Ferguson, N. M. Using *Wolbachia* for dengue control: insights from modelling. *Trends Parasitol.* **34**, 102–113 (2018).
46. Bhatt, S. et al. The global distribution and burden of dengue. *Nature* **496**, 504–507 (2013).
47. Salje, H. et al. Dengue diversity across spatial and temporal scales: local structure and the effect of host population size. *Science* **355**, 1302–1306 (2017).
This study uses genetic and epidemiological data to quantitatively characterize the spatial range and local population drivers of dengue virus transmission.
48. Reiner, R. C. Jr et al. Time-varying, serotype-specific force of infection of dengue virus. *Proc. Natl Acad. Sci. USA* **111**, E2694–E2702 (2014).
49. Andersson, N. et al. Evidence based community mobilization for dengue prevention in Nicaragua and Mexico (*Camino Verde*, the Green Way): cluster randomized controlled trial. *Br. Med. J.* **351**, h3267 (2015).
50. Cauchemez, S. et al. Association between Zika virus and microcephaly in French Polynesia, 2013–15: a retrospective study. *Lancet* **387**, 2125–2132 (2016).
51. Johansson, M. A., Mier-y-Teran-Romero, L., Reefhuis, J., Gilboa, S. M. & Hills, S. L. Zika and the risk of microcephaly. *N. Engl. J. Med.* **375**, 1–4 (2016).
52. Reiner, R. C. Jr et al. Quantifying the epidemiological impact of vector control on dengue. *PLoS Negl. Trop. Dis.* **10**, e0004588 (2016).
53. Ooi, E.-E., Goh, K.-T. & Gubler, D. J. Dengue prevention and 35 years of vector control in Singapore. *Emerg. Infect. Dis.* **12**, 887–893 (2006).
54. Fullman, N., Burstein, R., Lim, S. S., Medlin, C. & Gakidou, E. Nets, spray or both? The effectiveness of insecticide-treated nets and indoor residual spraying in reducing malaria morbidity and child mortality in sub-Saharan Africa. *Malar. J.* **12**, 62 (2013).
55. Lengeler, C. Insecticide-treated bed nets and curtains for preventing malaria. *Cochrane Database Syst. Rev.* **2**, CD000363 (2004).
56. Pluess, B., Tanser, F. C., Lengeler, C. & Sharp, B. L. Indoor residual spraying for preventing malaria. *Cochrane Database Syst. Rev.* **4**, CD006657 (2010).
57. West, P. A. et al. Indoor residual spraying in combination with insecticide-treated nets compared to insecticide-treated nets alone for protection against malaria: a cluster randomised trial in Tanzania. *PLoS Med.* **11**, e1001630 (2014).
58. Powell, J. R. & Tabachnick, W. J. History of domestication and spread of *Aedes aegypti*—a review. *Mem. Inst. Oswaldo Cruz* **108**, 11–17 (2013).
59. De Silva, P. M. & Marshall, J. M. Factors contributing to urban malaria transmission in sub-Saharan Africa: a systematic review. *J. Trop. Med.* **2012**, 819563 (2012).
60. WHO. Integrated vector management. http://www.who.int/neglected_diseases/vector_ecology/ivm_concept/en/. Accessed 17 February 2018.
61. Department of Control of Neglected Tropical Diseases. How to design vector control efficacy trials. Guidance on phase III vector control field trial design. Report No. WHO/HTM/NTD/VEM/2017.03 http://www.who.int/neglected_diseases/vector_ecology/resources/WHO_HTM_NTD_VEM_2017.03/en/ (World Health Organization, 2017).
62. Wolbers, M., Kleinschmidt, I., Simmons, C. P. & Donnelly, C. A. Considerations in the design of clinical trials to test novel entomological approaches to dengue control. *PLoS Negl. Trop. Dis.* **6**, e1937 (2012).
63. Snow, R. W., Guerra, C. A., Noor, A. M., Myint, H. Y. & Hay, S. I. The global distribution of clinical episodes of *Plasmodium falciparum* malaria. *Nature* **434**, 214–217 (2005).
64. White, M. T. et al. A combined analysis of immunogenicity, antibody kinetics and vaccine efficacy from phase 2 trials of the RTS,S malaria vaccine. *BMC Med.* **12**, 117 (2014).

© 2018 Springer Nature Limited. All rights reserved.

65. White, M. T. et al. Immunogenicity of the RTS,S/AS01 malaria vaccine and implications for duration of vaccine efficacy: secondary analysis of data from a phase 3 randomised controlled trial. *Lancet Infect. Dis.* **15**, 1450–1458 (2015). **This study analyses the data from the phase III trials of the RTS,S malaria vaccine to derive a mechanistic model of vaccine action and propose a predictive correlate of protection**.
66. Blackman, M. A., Kim, I.-J., Lin, J.-S. & Thomas, S. J. Challenges of vaccine development for Zika virus. *Viral Immunol.* **31**, 117–123 (2018).
67. Asher, J. et al. Preliminary results of models to predict areas in the Americas with increased likelihood of Zika virus transmission in 2017. Preprint at https://www.biorxiv.org/content/early/2017/09/18/187591 (2017).
68. WHO. Efficacy trials of ZIKV vaccines: endpoints, trial design, site selection. http://www.who.int/blueprint/what/norms-standards/ZIKV_final_report_public_version.pdf (2017).
69. Eisele, T. P. et al. Short-term impact of mass drug administration with dihydroartemisinin plus piperaquine on malaria in Southern Province Zambia: a cluster-randomized controlled trial. *J. Infect. Dis.* **214**, 1831–1839 (2016).
70. Brady, O. J. et al. Role of mass drug administration in elimination of *Plasmodium falciparum* malaria: a consensus modelling study. *Lancet Glob. Health* **5**, e680–e687 (2017).
71. Bigira, V. et al. Protective efficacy and safety of three antimalarial regimens for the prevention of malaria in young Ugandan children: a randomized controlled trial. *PLoS Med.* **11**, e1001689 (2014).
72. Capeding, M. R. et al. Clinical efficacy and safety of a novel tetravalent dengue vaccine in healthy children in Asia: a phase 3, randomised, observer-masked, placebo-controlled trial. *Lancet* **384**, 1358–1365 (2014).
73. Villar, L. et al. Efficacy of a tetravalent dengue vaccine in children in Latin America. *N. Engl. J. Med.* **372**, 113–123 (2015).
74. Winskill, P., Walker, P. G., Griffin, J. T. & Ghani, A. C. Modelling the cost-effectiveness of introducing the RTS,S malaria vaccine relative to scaling up other malaria interventions in sub-Saharan Africa. *BMJ Glob. Health* **2**, e000090 (2017).
75. World Health Organization. Malaria vaccine: WHO position paper—January 2016. *Wkly Epidemiol. Rec.* **91**, 33–51 (2016).
76. World Health Organization. Dengue vaccine: WHO position paper—July 2016. *Wkly Epidemiol. Rec.* **91**, 349–364 (2016).
77. WHO. Updated questions and answers related to information presented in the Sanofi Pasteur press release on 30 November 2017 with regards to the dengue vaccine Dengvaxia. http://www.who.int/immunization/diseases/dengue/q_and_a_dengue_vaccine_dengvaxia/en/ (2017).
78. Screaton, G. & Mongkolsapaya, J. Which dengue vaccine approach is the most promising, and should we be concerned about enhanced disease after vaccination?: The challenges of a dengue vaccine. *Cold Spring Harb. Perspect. Biol.* **10**, a029520 (2018).
79. WHO. Tables of malaria vaccine projects globally. http://www.who.int/immunization/research/development/Rainbow_tables/en/ (2017).
80. Barba-Spaeth, G. et al. Structural basis of potent Zika–dengue virus antibody cross-neutralization. *Nature* **536**, 48–53 (2016).
81. Fiorenzano, J. M., Koehler, P. G. & Xue, R.-D. Attractive toxic sugar bait (ATSB) for control of mosquitoes and its impact on non-target organisms: a review. *Int. J. Environ. Res. Public Health* **14**, 398 (2017).
82. Qualls, W. A. et al. Indoor use of attractive toxic sugar bait (ATSB) to effectively control malaria vectors in Mali, West Africa. *Malar. J.* **14**, 301 (2015).
83. Slater, H. C., Walker, P. G. T., Bousema, T., Okell, L. C. & Ghani, A. C. The potential impact of adding ivermectin to a mass treatment intervention to reduce malaria transmission: a modelling study. *J. Infect. Dis.* **210**, 1972–1980 (2014).
84. Werren, J. H., Baldo, L. & Clark, M. E. *Wolbachia*: master manipulators of invertebrate biology. *Nat. Rev. Microbiol.* **6**, 741–751 (2008).
85. Moreira, L. A. et al. A *Wolbachia* symbiont in *Aedes aegypti* limits infection with dengue, Chikungunya, and *Plasmodium*. *Cell* **139**, 1268–1278 (2009).
86. Walker, T. et al. The *w*Mel *Wolbachia* strain blocks dengue and invades caged *Aedes aegypti* populations. *Nature* **476**, 450–453 (2011).
87. Ferguson, N. M. et al. Modeling the impact on virus transmission of *Wolbachia*-mediated blocking of dengue virus infection of *Aedes aegypti*. *Sci. Transl. Med.* **7**, 279ra237 (2015). **This study combines experiemental work and modelling to estimate the likely impact of replacing wild-type A. aegypti with Wolbachia-infected mosquitoes on dengue disease transmission**.
88. Bian, G. et al. *Wolbachia* invades *Anopheles stephensi* populations and induces refractoriness to *Plasmodium* infection. *Science* **340**, 748–751 (2013).
89. Adelman, Z. N. *Genetic Control of Malaria and Dengue* (Academic, Amsterdam, 2015).
90. Gabrieli, P., Smidler, A. & Catteruccia, F. Engineering the control of mosquito-borne infectious diseases. *Genome Biol.* **15**, 535 (2014).
91. Adelman, Z. N. & Tu, Z. Control of mosquito-borne infectious diseases: sex and gene drive. *Trends Parasitol.* **32**, 219–229 (2016).
92. Marshall, J. M. & Taylor, C. E. Malaria control with transgenic mosquitoes. *PLoS Med.* **6**, e1000020 (2009).
93. Galizi, R. et al. A synthetic sex ratio distortion system for the control of the human malaria mosquito. *Nat. Commun.* **5**, 3977 (2014).
94. Burt, A. Site-specific selfish genes as tools for the control and genetic engineering of natural populations. *Proc. R. Soc. B* **270**, 921–928 (2003). **This study was one of the first to provide a rigorous population genetic assessment of the potential of using homing endonuclease genes to drive genetic modification of vector populations for disease control purposes**.
95. Deredec, A., Godfray, H. C. J. & Burt, A. Requirements for effective malaria control with homing endonuclease genes. *Proc. Natl Acad. Sci. USA* **108**, E874–E880 (2011).
96. Li, M. et al. Germline Cas9 expression yields highly efficient genome engineering in a major worldwide disease vector, *Aedes aegypti*. *Proc. Natl Acad. Sci. USA* **114**, E10540–E10549 (2017).
97. Gantz, V. M. et al. Highly efficient Cas9-mediated gene drive for population modification of the malaria vector mosquito *Anopheles stephensi*. *Proc. Natl Acad. Sci. USA* **112**, E6736–E6743 (2015).
98. Wang, S. & Jacobs-Lorena, M. Genetic approaches to interfere with malaria transmission by vector mosquitoes. *Trends Biotechnol.* **31**, 185–193 (2013).
99. Severson, D. W. & Behura, S. K. genome investigations of vector competence in *Aedes aegypti* to inform novel arbovirus disease control approaches. *Insects* **7**, 58 (2016).
100. Wesolowski, A. et al. Quantifying the impact of human mobility on malaria. *Science* **338**, 267–270 (2012).

Acknowledgements I thank the Medical Research Council, the National Institute of Health Research Health Protection Research Unit programme, the National Institute of General Medical Sciences 'MIDAS' programme and the Bill and Melinda Gates Foundation for research funding. Among many others, A. Ghani, A. Burt and S. O'Neill have informed my understanding of this topic during many illuminating discussions.

Competing interests The author declares no competing interests.

Additional information

Reprints and permissions information is available at http://www.nature.com/reprints.

Correspondence and requests for materials should be addressed to N.M.F.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

REVIEW

https://doi.org/10.1038/s41586-018-0327-4

Challenges and recent progress in drug discovery for tropical diseases

Manu De Rycker[1], Beatriz Baragaña[1], Suzanne L. Duce[2] & Ian H. Gilbert[1]*

Infectious tropical diseases have a huge effect in terms of mortality and morbidity, and impose a heavy economic burden on affected countries. These diseases predominantly affect the world's poorest people. Currently available drugs are inadequate for the majority of these diseases, and there is an urgent need for new treatments. This Review discusses some of the challenges involved in developing new drugs to treat these diseases and highlights recent progress. While there have been notable successes, there is still a long way to go.

Infectious tropical diseases mainly affect low- and middle-income countries (LMICs). Symptoms include impaired cognitive and physical development in children (Box 1), complications in pregnancy, fever, nausea, diarrhoea, dehydration, anaemia, rashes, lesions, deformities, blindness, organ failure, haemorrhage, neurological problems, seizures and coma, and, in many cases, death. The WHO (World Health Organization) estimated that the diseases listed in Fig. 1 and Table 1 were responsible for more than 4 million deaths in 2015[1] and the loss of more than 250 million disability-adjusted life years (DALYs) owing to ill-health, premature disability or early death[2].

Infectious tropical diseases include those defined by the WHO as neglected tropical diseases[3] (NTDs). In addition, diseases such as malaria, tuberculosis, HIV/AIDS, multi-drug-resistant Gram-negative bacterial infections, diarrhoea (from a variety of pathogens) and hepatitis disproportionately affect tropical countries. The most common tropical diseases, grouped by causative agent, include: (i) viruses: Ebola, HIV/AIDS, Lassa fever, Marburg virus disease, Rift Valley fever, yellow fever, Zika disease, and the NTDs chikungunya, dengue fever and rabies; (ii) bacteria: bubonic plague, shigellosis, tuberculosis, typhoid fever, typhus, disease caused by drug-resistant Gram-negative bacteria, and the NTDs buruli ulcer, leprosy, mycetoma, trachoma and yaws; (iii) protozoa: cryptosporidiosis, malaria, and the NTDs Chagas disease, human African trypanosomiasis (HAT) and leishmaniases (visceral, cutaneous, post-kala-azar dermal and mucocutaneous); (iv) fungi: cryptococcosis and the NTD mycetoma; (v) helminths or metazoan worms: the NTDs cysticercosis, dracunculiasis (Guinea worm), echinococcosis, foodbourne trematodiases (clonorchiasis, opisthorchiasis, fasciolosis), lymphatic filariasis, onchocerciasis, schistosomiasis and soil-transmitted helminthiases (ascariasis, hookworm, trichuriasis, strongyloidiasis). Co-infection with these diseases is commonplace, and can exacerbate symptoms and complicate treatments (Box 2).

Infectious tropical diseases inflict a heavy economic burden, costing LMIC economies billions of dollars every year. For example, the annual global cost of dengue in 2013 was estimated at US$8.9 billion[4], and a study carried out in the Philippines indicated that schistosomiasis resulted in the loss of an average of 45.4 work days per infected person per year[5].

For many of these diseases the current treatments are unsatisfactory, and for some there are few or no treatment options available. There is an urgent need for new, safe, low-cost and efficacious treatments. Tackling these diseases effectively requires an integrated multi-sectoral approach that links chemotherapy with prevention initiatives such as improved water supply, sanitation and hygiene, vaccines and vector control. Various governmental and non-governmental organizations, charities, foundations, product-development partnerships, academic groups and pharmaceutical companies are supporting drug discovery for tropical diseases. A breakdown of funding is summarized in the G-Finder[6] and IFPMA 2017[7] reports. In 2016 about two-thirds of the total funding for research and development for neglected diseases was spent on HIV/AIDS (34%), malaria (18%) and tuberculosis (18%)[6]. Overall research and development funding has markedly increased in the last decade but the gains have not been distributed evenly across diseases. Greater efforts, resources and funding are urgently needed to meet the considerable unmet medical needs.

In this Review, we will consider the specific challenges associated with drug discovery for tropical diseases and examine some of the recent successes.

Discovery pathways and challenges

Figure 2 shows the pathway for a typical small-molecule drug-discovery programme for an infectious disease.

The drug–discovery pathway

'Hit discovery' is the first step on the drug-discovery pathway, and is where chemical starting points are identified. This usually involves screening collections of compounds against a molecular target, typically an enzyme (target-based screening), or against whole organisms (cell-based or phenotypic screening). Alternative approaches in hit discovery include structure-based drug discovery, re-purposing of drugs from other disease areas, and in silico methods. Next, 'hit expansion' is performed to verify if a hit is genuine and whether there is scope for further work with the chemotype. The compound is then refined through a cyclic process of 'design–make–test' in the 'hits to lead' phase, until it has significant activity, typically in an animal model of infection. Subsequently, the biological activity, pharmacokinetics and safety profile of the series are optimized and balanced during the 'lead optimization' phase. 'Candidate selection' marks an important milestone in the drug-discovery pathway, as it involves selecting a single compound for progression. In this process, regulatory toxicology and scale-up are carried out to enable initial human studies. In human 'clinical trials', the compound is first tested in healthy volunteers to measure pharmacokinetics and safety (phase I); this is followed by

[1]Drug Discovery Unit, Wellcome Centre for Anti-Infectives Research, Division of Biological Chemistry and Drug Discovery, School of Life Sciences, University of Dundee, Dundee, UK. [2]Medicines Monitoring Unit (MEMO), Division of Molecular and Clinical Medicine, School of Medicine, University of Dundee, Dundee, UK. *e-mail: i.h.gilbert@dundee.ac.uk

© 2018 Springer Nature Limited. All rights reserved.

Box 1

Infants and children

Many of the infectious tropical diseases disproportionately affect infants and children. In the case of malaria, the majority of deaths occur in children under the age of five, as they are immune-naive. Diarrhoea is a major problem among children, causing hundreds of thousands of deaths per year. The Global Enteric Multicentre Study (GEMS), a case–control study conducted at seven sites in Africa and south Asia, is the most comprehensive study of childhood diarrhoea to date[112,113]. The most common pathogens found in this study were *Shigella* spp., rotavirus, adenovirus 40/41, heat-stable enterotoxin-producing *Escherichia coli*, *Cryptosporidium* spp. and *Campylobacter* spp. In some cases, mixed infections were an issue. Some tropical diseases, such as cryptosporidiosis, schistosomiasis and soil-transmitted helminths[114,115], are associated with malnutrition, growth stunting and impaired cognitive development in children. In addition, the drug-development pathway for infants and children is complicated. Pharmacokinetics are much more difficult to predict in infants and children and additional safety testing is required, making clinical trials challenging. Furthermore, ethical regulations in clinical trials for children are more complex.

studies to establish efficacy (phase II) and then large-scale efficacy and safety studies (phase III).

Target product profiles and compound progression criteria

The drug-discovery pathway should be guided by 'target product profiles' (TPPs). These describe the desired features required of the final drug product, such as duration of treatment, whether the treatment is oral or parenteral, cost of treatment, acceptable safety margins, among others[8]. TPPs have been defined for malaria[9], leishmaniasis[10], HAT[10], Chagas disease[11], cryptosporidiosis[12] and dengue fever[13], but are absent for many other diseases, hampering the drug-discovery process. TPPs are used to develop 'compound progression criteria', which inform the required disease-specific profile of a compound at every step along the drug-discovery process. These include, for example, activities in particular assays, pharmacokinetic profiles, and selectivity parameters[8]. Medicines for Malaria Venture (MMV) has several examples of compound progression criteria (https://www.mmv.org/research-development/information-scientists).

Challenges

For many tropical infectious diseases, there is little or no precedent for developing small-molecule drugs. This is exacerbated by insufficient understanding of the pathogen biology. As a result, there are often no relevant cellular or predictive animal models of the human disease. There is also frequently little or no data from clinically active molecules to help researchers to define the profile that a new drug should have in terms of its activity in preclinical assays and animal models, and its pharmacokinetics. The key issues facing researchers seeking to discover new drugs to treat tropical diseases are summarized in Fig. 2 and are considered below.

Biological challenges. There are very few well-validated molecular drug targets for tropical diseases, in part owing to a lack of understanding of the detailed biology of many of the pathogens. For example, the functions of many proteins are unknown, or have only been inferred from other organisms. Therefore a key challenge is the selection of a suitable molecular target to limit the risk of subsequent failure[8]. Criteria used to help to select appropriate targets include essentiality, druggability, assayability and the opportunity for selectivity over host orthologues[8]. Even if a target complies with all these criteria, inhibitors frequently fail to demonstrate cell-based activity owing to poor permeability, inability to compete with high substrate concentrations in cells, or because the compounds cannot inhibit the target sufficiently to kill the pathogen. The deficiency of genetic tools for many disease-relevant organisms is a key reason why so little essentiality data are available. However, new technologies are emerging[14–16], including CRISPR–Cas9[17,18], which offer the prospect of markedly increasing the number of validated targets available in the near future.

Large-scale cell-based (phenotypic) screening is currently a popular approach to identify new chemical start points, as it does not require prior knowledge of the molecular target[19,20]. Developing assays that have sufficient throughput to allow screening of compound libraries while maintaining relevance to disease pathophysiology is challenging, but helps to reduce attrition when using in vivo models and, ultimately, in clinical trials[21]. As assays increase in relevance, they usually become more complex and require more resources, both in terms of specialist equipment as well as staff to develop and run the assays. Other challenges may include the lack of robust in vitro culture systems (for example, for *Cryptosporidium*), the inherent safety challenges of handling disease-causing agents in bulk, the lack of standardization between different laboratories working on the same organisms, and the use of laboratory strains and cell lines rather than recent clinical isolates and primary cells[21,22].

Existing animal models of tropical disease infections are often poorly representative of the human disease. For example in onchocerciasis (river blindness), the causative worm (*Onchocerca volvulus*) is only infective to primates; consequently, related, but different, worm species are used in mice or cattle models. *Plasmodium falciparum*, the pathogen that causes the most severe form of human malaria, does not infect rodents normally; however, a SCID mouse model has been developed that can be infected by *P. falciparum*[23]. Work is also ongoing to develop a mouse model of human malaria caused by *Plasmodium vivax*[24]. Most mouse models of tuberculosis do not replicate the human disease well, and do not form the granulomas that are typical of the human disease. Encouragingly, progress is being made on developing animal models that more closely replicate human pathology[25–27].

Recent human clinical trials of azoles, which inhibit the enzyme CYP51, for treatment of Chagas disease have not been successful. A mouse model has now been developed that can distinguish between benznidazole, which is known to have clinical efficacy, and an azole (posaconazole), which does not[28]. Better differentiation at this preclinical stage should improve the success rates of clinical trials.

Quiescence and dormancy. Dormant infections further complicate the aetiology of some diseases. In malaria, *P. vivax* and *Plasmodium ovale* can remain dormant (as hypnozoites) in hepatocytes of the human host for weeks to years, potentially causing relapse[29]. Latent tuberculosis, in which clinical symptoms are absent but there is a risk of progression to clinical disease, is estimated to affect a third of the global population[30]. Dormancy has recently been identified in *Trypanosoma cruzi*[31]. Quiescent and dormant pathogens have reduced metabolism compared to actively dividing pathogens and may therefore be less susceptible to drugs. Developing assays for these forms of pathogens is problematic. To effectively treat and eliminate these diseases, new drugs are needed that effectively tackle the reservoir of dormant and quiescent infections.

Chemistry. It is essential that the compound used for treatment can be dosed appropriately, usually orally, and can reach the part of the body where the pathogen is located. The physicochemical properties of a molecule, such as molecular weight, solubility, lipophilicity, charge, and number of hydrogen bond donors and acceptors are fundamental to achieving this.

In some cases, penetration into the pathogen is problematic. For example, Gram-negative bacteria are surrounded by an outer membrane; drug molecules can penetrate this through proteins that form channels through the membrane called porins, which tend to select for hydrophilic molecules. The drug then has to traverse the cell membrane, which is a lipid bilayer; this requires compounds with

© 2018 Springer Nature Limited. All rights reserved.

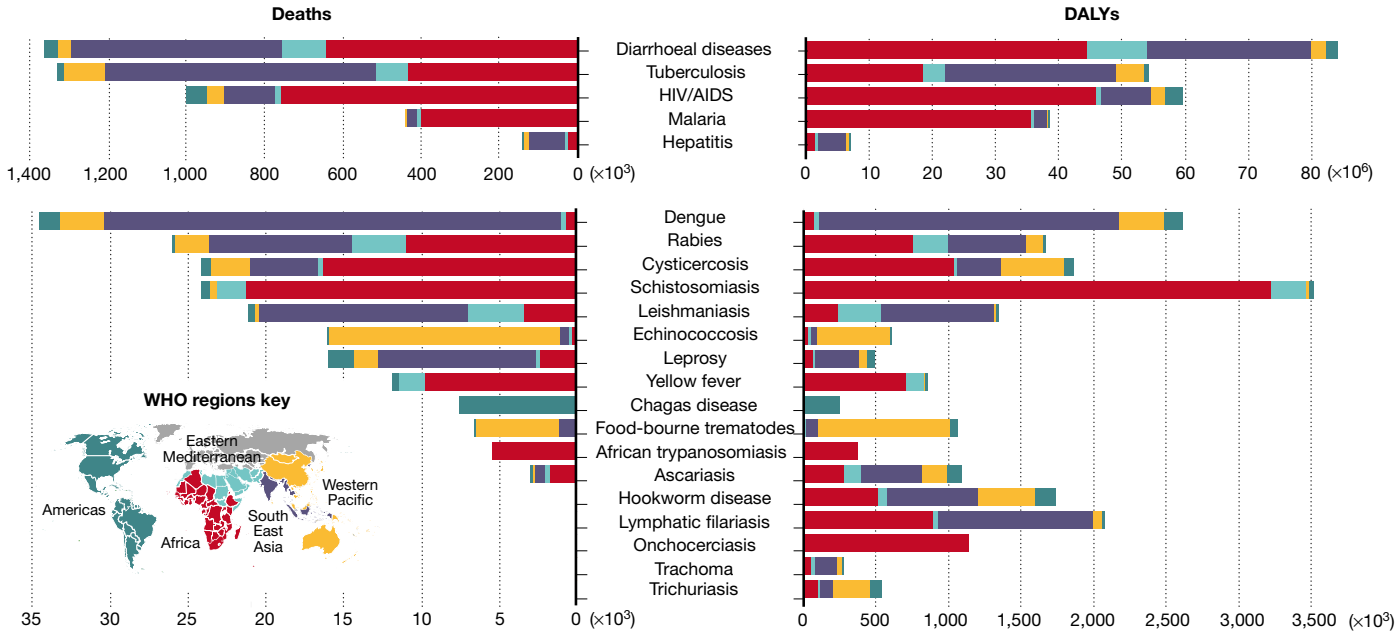


Fig. 1 | Deaths and disability-adjusted life years (DALYS) due to tropical diseases. From WHO estimates of cause-specific mortality (2015)[1] and disease burden (2015)[2]. Diarrhoeal diseases, tuberculosis, HIV/AIDS, malaria and hepatitis are shown separately as they have a much larger impact. Data are shown by WHO region for regions that contain tropical countries.

a degree of lipophilicity. Therefore, to penetrate into the cytoplasm, compounds require very defined chemical properties, and are likely to be more polar and charged than typical drugs[32,33]. Having penetrated to the cytoplasm, the drug then risks being pumped out of the bacteria through efflux transporters. Pathogens such as *T. cruzi, Chlamydia trachomatis, Mycobacterium tuberculosis* and viruses spend the majority of their life cycle inside human host cells, which can impede drug access. Some pathogens localize to extreme intracellular environments such as acidic organelles (for example, *Leishmania* spp. or *Salmonella* spp.) or necrotic granulomas (for example, *M. tuberculosis*). In the latter case, it is a major challenge to design drugs that can penetrate through the non-vascularized lipid-rich caseum to access the pathogen[34]. Other infections are localized at sites that are protected by a blood–tissue barrier, such as the central nervous system (for example, HAT, *Cryptococcus neoformans*, *Taenia solium*, or encephalitis- or meningitis-causing viruses or bacteria) or the eyes (for example, *C. trachomatis*). To reach these sites, drugs need to pass through endothelial cells without being pumped back into the blood by P-glycoprotein[35].

The nature of these diseases means that the cost of treatment must be low. For malaria, the aim is to produce a treatment that costs less than US$1. This implies a short and cheap chemical synthesis and precludes complex formulations to counter issues such as poor solubility. Owing to the lack of temperature-controlled supply chains in many regions, compounds that are stable over long periods of time at high temperatures (>40 °C) and high humidity are required. Finally, patient compliance is also challenging, and is often compounded by minimal medical support. Consequently, it is essential to minimize the number of doses and complexity of dosing regimens. For example, MMV is aiming for a single-dose treatment for bloodstream malaria[9].

Resistance. Drug resistance is a considerable problem for treatment of most infectious diseases, in particular for viral, bacterial and apicomplexan infections[36–38]. Poor management of treatments can accelerate the occurrence of resistance. The most publicized current example is antibacterial drug resistance. In humans, antibiotics are sometimes taken when not required, or given at sub-therapeutic doses, potentially increasing the proportion of resistant bacteria within a population. This problem is aggravated by the large amounts of antibiotics in the environment owing to indiscriminate use in agriculture and aquaculture.

Table 1 | The pathogens of common tropical diseases

Disease	Pathogen	Type
Diarrhoeal diseases	Multiple pathogens	Bacteria, Viruses, Protozoa
Tuberculosis	*Mycobacterium tuberculosis*	Mycobacterium
HIV/AIDS	Human immunodeficiency virus	Virus
Malaria	*Plasmodium* spp.	Protozoan
Hepatitis	Hepatitis A, B, C, D, E	Virus
Dengue	Dengue virus	Virus
Rabies	Rhabdoviridae	Virus
Cysticercosis	*Taenia solium*	Tapeworm
Schistosomiasis	*Schistosoma* spp.	Helminth
Leishmaniasis	*Leishmania* spp.	Protozoan
Echinococcosis	*Echinococcus* spp.	Tapeworm
Leprosy	*Mycobacterium leprae*	Mycobacterium
Yellow fever	Yellow fever virus	Virus
Chagas diseases	*Trypanosoma cruzi*	Protozoan
Food-bourne trematode	*Clonorchiasis, Opisthorhiasis, Fascioliasis, Paragonimisaisis*	Helminth
African trypanosomiasis	*Trypanosoma brucei*	Protozoan
Ascariasis	*Ascaris lumbricoides*	Soil-transmitted helminth
Hookworm diseases	*Ancylostoma duodenale, Necator americanus*	Soil-transmitted helminth
Lymphatic filariasis	*Wuchereria bancrofti, Brugia malayi, Brugia timori*	Nematode
Onchocerciasis	*Onchocerca volvulus*	Nematode
Trachoma	*Chlamydia trachomatis*	Bacterium
Trichuriasis	*Trichuris trichiura*	Soil-transmitted helminth

The diseases referred to in Fig. 1 are described with their corresponding pathogens.

© 2018 Springer Nature Limited. All rights reserved.

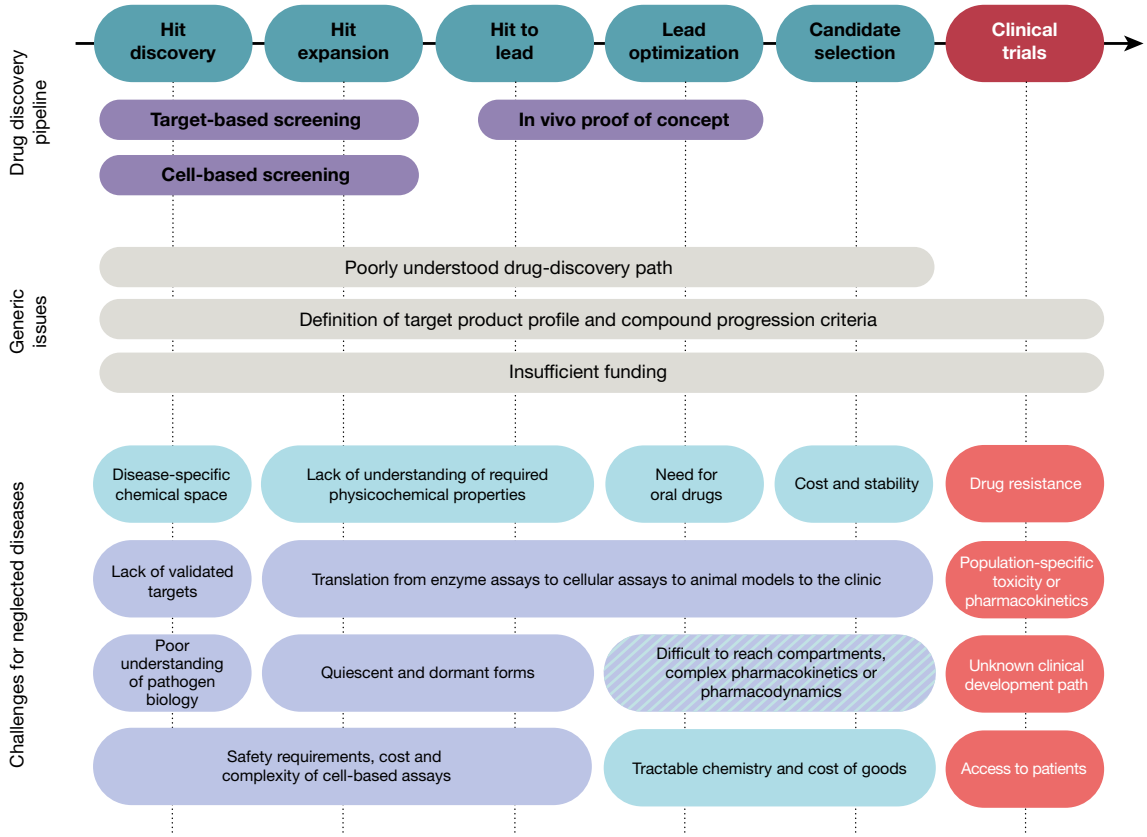


Fig. 2 | The drug-discovery process. Top, the stages of the drug-discovery process; middle, generic challenges; bottom, additional key challenges at particular stages of the drug-discovery process for neglected diseases. Biological and pharmacokinetic issues are shown in lilac, chemical issues are shown in light blue and clinical issues are shown in red.

Combination therapy is being adopted for many diseases in an attempt to slow development of drug resistance. This has been particularly successful in the treatment of HIV/AIDS, in which the resistance rate for single therapy[39] is exceptionally high. Combination treatment has also proved very effective against tuberculosis and malaria. Worryingly, in the case of malaria, resistance to artemisinin combination therapy, the principal form of treatment[38], is now developing in southeast Asia. There is therefore a need for drugs with novel modes of action to deliver new combination therapies for which there is no clinical resistance to any of the components.

Progress

Despite all these challenges, progress is being gradually made. Some examples are highlighted below, although this is not a comprehensive list. Such successes are due to coordinated and substantial investments in research, along with concerted efforts to understand the TPPs and drug-discovery pathways.

Apicomplexans

With 200 million cases a year, the human and economic cost of malaria is extremely high[40]. Africa bears a disproportionately high share of the burden with more than 400,000 deaths a year. Multiple new types of antimalarial drugs are needed to overcome resistance, give single-dose treatments, prevent relapse of *P. vivax* infections, block transmission and to act as chemopreventatives. The malaria drug-discovery portfolio has markedly improved over the past decade[41]. The starting points for the majority of the compounds currently in development were identified by phenotypic screening in collaborations with MMV. In many cases, target deconvolution has subsequently led to the identification of the molecular targets. There are a number of novel compounds in clinical trials supported by or in collaboration with MMV (Fig. 3). The following are derived from phenotypic screens and are in phase II trials: the spiroindolone KAE609, the imidazolopiperazine KAF156[42] (a combination trial with lumefantrine), the synthetic trioxolane artefenomel (OZ439, in combination with ferroquine)[43] and the 2-aminopyridine MMV048. The mode of action has been determined for some of these (*P. falciparum* (*Pf*) ATP4 for KAE609[44] and *Pf*PI4K for MMV048[45]), whereas the modes of action of KAF156 and OZ439 remain uncertain. KAF156 has been linked to the cyclic amine resistance locus (Pf*carl*)[46] and OZ439 probably acts through formation of carbon radicals, whereas ferroquine acts to prevent haem detoxification. In phase I trials are: the dihydroisoquinoline SJ733, another *Pf*ATP4 inhibitor[47]; and M5717, which is active against translation elongation factor 2[48]. Two compounds derived from target-based projects are also in clinical trials: the dihydroorotate dehydrogenase inhibitor DSM265[49] is in phase II trials, and the dihydrofolate reductase inhibitor P218 is in phase I trials.

We envisage an increase in target-based drug discovery as mode-of-action studies are revealing the chemically validated targets of many of the recently identified phenotypic hits. There are systematic efforts trying to identify the mode of action of phenotypic hits[50]. New types of phenotypic screens have been developed that target life-cycle stages in addition to, or other than, the asexual blood stages[51,52]. Clinical development for malaria has accelerated thanks to the development of human challenge models in which healthy volunteers are infected with treatable malaria to obtain an early indication of drug efficacy[53,54].

Cryptosporidiosis is a major cause of diarrhoea in children in tropical countries (Box 1). Nitazoxanide, the current treatment for the disease, has variable efficacy in immunocompetent patients and is not effective in immunocompromized patients[55,56]. There is an urgent need for new drugs to treat children under 24 months of age, especially those who are malnourished, immunocompromized or suffering from chronic diarrhoea.

There have been recent advances in the cryptosporidiosis discovery pipeline. A series of bumped kinase inhibitors of *Cryptosporidium parvum* calcium-dependent kinase 1[57] has been developed, an example of which is BKI1534. Promisingly, recent advances in *C. parvum*

© 2018 Springer Nature Limited. All rights reserved.

Box 2

Co-infections

Co-infections are associated with high prevalence and the extensive geographical overlap of some of these diseases. A person who is already infected with one pathogen may, in some cases, have an increased risk of infection with another pathogen[114]. Some of the issues around co-infection are:

- Co-infections frequently accelerate disease progression and outcomes as one disease enhances the effect of another. For example, malaria infection increases the viral load of HIV-infected patients, also making it easier for that person to pass on the HIV virus[116]. HIV–tuberculosis co-infection is particularly problematic; HIV weakens the immune system, allowing latent tuberculosis to become activated and progress more rapidly. In turn, tuberculosis also accelerates the progression of HIV infection[117].
- Co-infections can result in complications in pregnancy and impair growth and development in children. For example, co-infections of malaria and hookworm can lead to severe depletions in haemoglobin resulting in severe anaemia[118].
- Co-infections can complicate treatments due to drug–drug interactions which can affect efficacy and can give rise to cumulative drug toxicities producing adverse side effects[119]. This is a particular problem in HIV–tuberculosis co-infection. Rifampicin, which is commonly used to treat tuberculosis, causes induction of cytochrome P450s. The latter metabolize some of the more commonly used anti-HIV drugs[120]. This means that careful selection of treatments is required.
- Co-infections can affect how diseases are treated. For example, ivermectin treatment for onchocerciasis may induce severe adverse reactions in some people with severe *Loa loa* infection[121,122].

high-content imaging infection assays in human intestinal epithelial cells have made phenotypic screening of medium- and large-size compound libraries possible. A screen of compounds active against malaria has led to the identification of a preclinical candidate, KDU731[58], an inhibitor of the *Cryptosporidium* phosphatidylinositol 4-kinase. Clofazimine, an approved drug for the treatment of leprosy, was identified from a screen of bioactive compounds[59] and is entering clinical trials for cryptosporidiosis.

Kinetoplastid diseases

This area has recently been reviewed in detail[60]. Mortality rates for HAT (also known as sleeping sickness) have decreased substantially in recent years[61]; however, there is still a considerable disease burden in central Africa. The advent of nifurtimox–eflornithine (NECT) combination therapy in 2009 marked an important step in the treatment of stage two *T. brucei gambiense* HAT. This has reduced the duration of treatment as well as decreasing adverse effects[62]. Nevertheless, treatment is still problematic because the two subspecies that cause the disease (*T. brucei gambiense* and *T. brucei rhodesiense*) and the two stages (acute and central nervous system) require different drugs[61]. The requirement for staging patients before treatment also limits uptake, as it involves a painful and technically challenging lumbar puncture[63]. The ongoing development of two new drugs has the potential to transform the clinical management of HAT[64]. The first of these drugs is fexinidazole, a nitroaromatic compound developed in the 1970s and 1980s as a broad-spectrum antibiotic, which was more recently found to have a suitable profile for clinical development for HAT[65]. Drugs for Neglected Diseases initiative (DNDi) completed a phase III clinical trial in 2017 and reported non-inferiority over NECT for both stages of the disease with a ten-day oral treatment[66]. The second is an oxaborole compound, SCYX-7158 (acoziborole)[67], which is currently undergoing a pivotal phase II/III trial[10]. This has the potential to provide a single-dose oral treatment for all forms of HAT.

Less progress has been made for leishmaniasis and Chagas disease. Current treatments, particularly liposomal amphotericin-B (which is given intravenously), together with other measures, such as vector control and effective surveillance, have led to progress towards elimination or control of visceral leishmaniasis (VL)[68] in the Indian subcontinent. However, current drugs are less effective or unsuitable in other areas of the world. New oral drugs are urgently needed to treat VL in these regions and to provide improved treatment options in the Indian subcontinent. Furthermore, HIV–VL co-infection is also currently poorly treated. Several potential oral treatments identified through phenotypic screening are in preclinical development, including an oxaborole and a nitroimidazole[69]. The discovery of a pan-kinetoplastid proteasome inhibitor[70] (GNF6702, Fig. 3) offers potential for development of new drugs for leishmaniasis, Chagas disease and HAT. The majority of preclinical work on leishmaniasis focuses on VL, the deadly form of the disease, rather than cutaneous leishmaniasis (CL). New oral therapies for CL are urgently needed as well. Research on Chagas disease suffered a setback recently as a consequence of the clinical failures of the CYP51 inhibitors posaconazole and fosravuconazole[71,72]. Currently, few treatments are under development for this challenging disease.

Viruses

Viruses such as HIV, hepatitis C virus (HCV), dengue and rabies have a large global impact (Fig. 1), resulting in over 1 million deaths in 2015[1]. Effective treatments for HIV and HCV have been developed[73,74]. Other viruses, including flaviviruses, arenaviruses, coronaviruses and filoviruses still pose a considerable threat for tropical countries and others, and pose a particular risk of global pandemics, as exemplified by the recent outbreaks of Zika[75], chikungunya[76], dengue[77], Ebola[78] and Middle East respiratory syndrome (MERS)[79]. Since future epidemics are almost inevitable, and could be triggered by any member of these families, drug-discovery efforts rightly focus on broad-acting antivirals[80]. These diseases have received relatively little attention in terms of small-molecule drug discovery. However, both target-based and cell-based hit discovery approaches have yielded new compounds of interest, but to date none of these have been developed to a similar degree as the HCV and HIV drugs[80]. The most advanced compounds are favipiravir, GS-5734 (remdesivir) and BCX4430 (galidesivir), which target the viral polymerase (Fig. 3). Favipiravir was tested in humans during the Ebola virus epidemic of 2014. Although conditions complicated the study, it was concluded that further investigation was warranted[81]. GS-5734 has shown activity in rhesus monkey model of Ebola[82], and was also used in humans during the 2014 Ebola outbreak. It is currently undergoing a phase II clinical trial in West Africa. Both favipiravir and GS-5734 are being considered for use in the 2018 Ebola virus outbreak in the Democratic Republic of the Congo[83]. BCX4430 has demonstrated complete protection against Marburg virus in non-human primates and against Ebola in rodent models when administered within 48 h of infection. This compound also shows broad activity against other viruses, including arenaviruses and flaviviruses[84,85], and is currently in clinical development. Inhibitors that target the viral protease are not as far advanced (with the exception of HIV and HCV); however, they also show promise as broad-spectrum antivirals. The potential of targeting host pathways is also being explored[80,86]. Although new small-molecule antivirals are being discovered, there is an urgent need to translate more such treatments into the clinic, alongside ongoing vaccine development.

Bacterial diseases

Tuberculosis continues to be a major problem across the world. As with malaria, combination treatments are used against tuberculosis to combat resistance, and can circumvent issues of tuberculosis residing in multiple environments and metabolic states. The need for combination therapy complicates the progression of new compounds through clinical trials (https://www.tballiance.org/portfolio). Recently, two new drugs, bedaquiline and delamanid, have been conditionally approved for the treatment of specific levels of drug-resistant tuberculosis on

© 2018 Springer Nature Limited. All rights reserved.

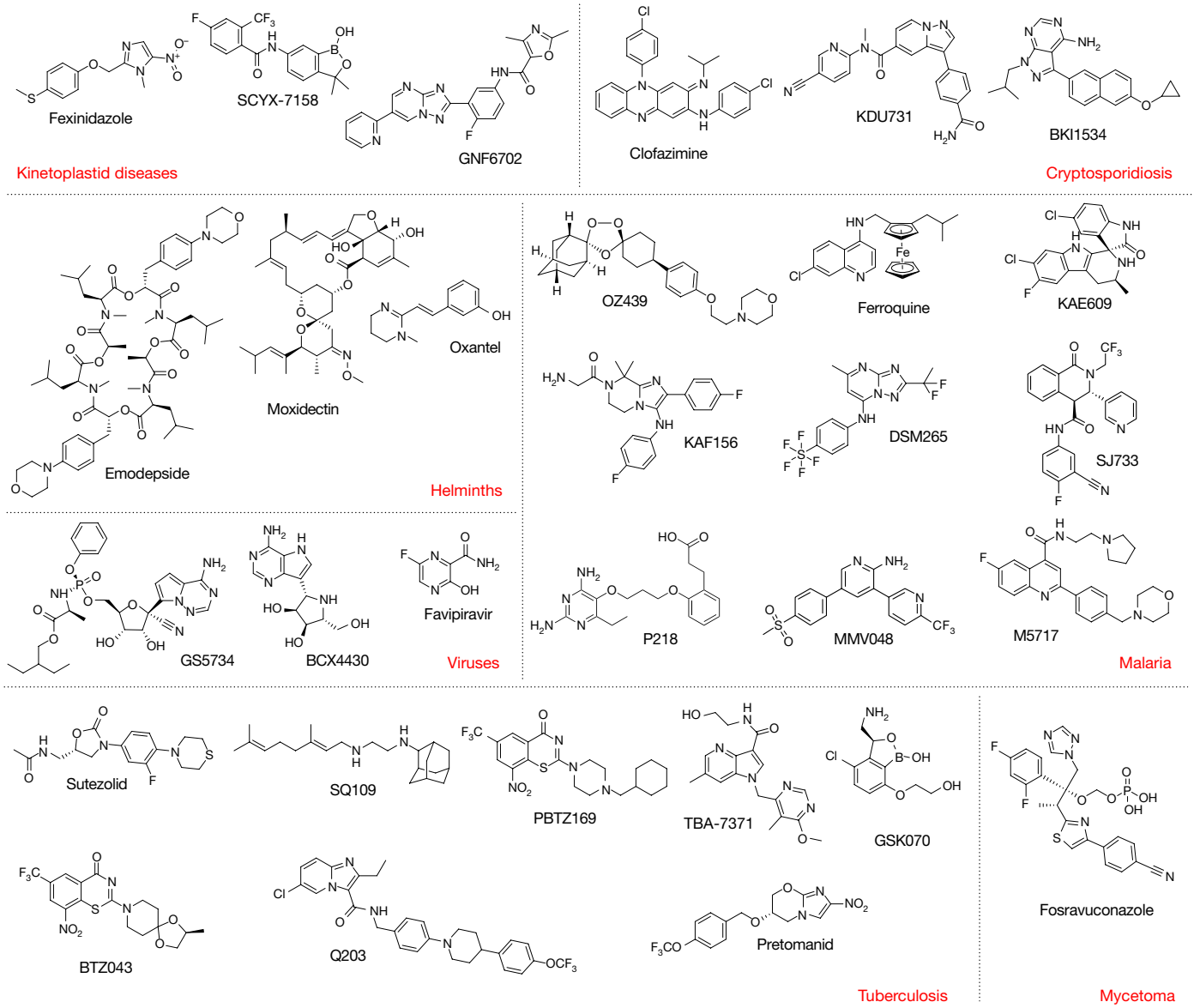


Fig. 3 | Key compounds discussed in the text.

the basis of phase IIb trials, and are currently undergoing phase III studies[87]. Bedaquiline targets ATP synthase[88] and is active against the latent forms of the bacterium[89]. Delamanid is a nitro heterocycle; its precise mode of action has not been determined, but it is probably a prodrug that requires activation inside the bacterium. It has been implicated in inhibition of mycolic acid biosynthesis, which is critical for the structure of the cell wall of *M. tuberculosis*[90,91].

A number of other compounds that are currently under development for treatment of tuberculosis target protein synthesis, the respiratory chain or the cell envelope (https://www.tballiance.org/portfolio and https://www.newtbdrugs.org/pipeline/clinical). Several different steps in protein synthesis are being studied. Linezolid has demonstrated activity against drug-resistant tuberculosis, and has a marked effect on patients with extensively drug-resistant tuberculosis[92]. It is a member of the oxazolidinone class of molecules, and targets the 50S ribosomal subunit. However adverse effects are associated with the extended treatment periods with linezolid that are required for multi-drug-resistant tuberculosis. This is linked to mitochrondrial toxicity[123]. Other oxazolidinones, such as sutezolid and delpazolid, are undergoing clinical trials[87]. Avoiding mitochondrial toxicity should allow development of oxazolidinone analogues with higher tolerability. GSK070, an oxaborole inhibitor of leucyl-tRNA synthetase, an enzyme involved in protein synthesis, is progressing through clinical trials[93]. GSK070 forms a covalent adduct with the 3′-adenosyl acceptor nucleotide of the tRNA, its boron binding to the 2′- and 3′-hydroxyl groups. This complex is formed in the editing site, inactivating the enzyme.

Q203 has just entered clinical trials and targets the respiratory pathway[88], inhibiting cytochrome bc_1. This is on the same pathway that is targeted by bedaquiline, although it acts at a different step. Hitting multiple points on this pathway appears to enhance killing of the bacterium[88].

β-lactams, which target cell wall biosynthesis, are also being re-investigated for treatment of tuberculosis. Carbapenems appear to have the most potential in this class of compounds, because of the proteins that they target and because they are much weaker substrates of the *M. tuberculosis* β-lactamases[87]. Several enzymes involved in cell envelope biosynthesis are promiscuous targets, including DprE1 and MmpL3. The former is involved in the biosynthesis of an arabinogalactan polysaccharide that is critical for cell wall biosynthesis. Benzothiazones have been found to be covalent inhibitors of this enzyme; one of these compounds, BTZ043, is in preclinical development, and another, PTBZ169, is undergoing clinical evaluation (https://www.newtbdrugs.org/pipeline/clinical). Two non-covalent inhibitors

© 2018 Springer Nature Limited. All rights reserved.

of DprE1 are also undergoing clinical evaluation, both in phase I, an azaindole, TBA-7371, and OPC-167832. MmpL3 is required for the export of mycolic acid precursors, which are essential for cell wall biosynthesis. Compound SQ109, which targets this enzyme and probably others[87], is in clinical trials[94]. Several preclinical drug-discovery programmes are targeting the enoyl reductase InhA, which is involved in the biosynthesis of mycolic acid. InhA is also the target of the prodrug isoniazid.

The nitroimidazole pretomanid (PA-824), which is active against both replicating and non-replicating mycobacteria, is in clinical trials. Like delamanid, pretomanid is likely to be a prodrug. A number of other compounds[124] are undergoing clinical evaluation (https://www.newtbdrugs.org/pipeline/clinical), and several fluoroquinolines are being investigated in combination therapies, including moxifloxacin and levofloxacin.

Whereas antibacterial drug resistance is not technically a neglected or tropical disease, its impact is most clearly seen in LMICs, where there are very high levels of drug resistance. The WHO has recently published a priority list of pathogens[95]. The priority 1 pathogens are: *Enterobacteriaceae*, which are resistant to carbapenems and third generation cephalosporins, and *Acinetobacter baumannii* and *Pseudomonas aeruginosa*, which are carbapenem-resistant. New drugs are urgently needed to tackle these pathogens, but there are few drugs in clinical development that address them[96]. In addition, several bacterial infections, such as shigellosis and typhoid, almost exclusively affect LMICs. Drug resistance is a major issue, and worryingly, in some instances, the resistant bacteria have a higher fitness than wild-type bacteria[97,98]. Treatment is reliant on existing antibacterials.

Helminths

At least one billion people, and probably more, are estimated to have one or more helminth infections[99]. Nevertheless, the only new antihelminthic drug to emerge in the last 30 years is tribendimidine, which shows promising broad-spectrum activity. Recent clinical trials show that it is a suitable alternative for the benzimidazoles and praziquantel in various helminth infections[100–102]. Beyond this, development of new treatments is focused on repurposing drugs. This has mostly involved anthelminthics that are approved for veterinary use. Oxantel pamoate, a licensed veterinary drug since 1974, has shown promising efficacy in human clinical trials for trichuriasis (whipworm), one of the more difficult-to-treat soil-transmitted helminths[103], and in combination with albendazole and pyrantel pamoate, it has recently demonstrated high clinical efficacy against hookworm infections[104]. Two other veterinary drugs, emodepside and moxidectin, show promising activity against several human helminths, and are currently undergoing clinical trials for onchocerciasis[69,105]. Some antimalarial compounds show activity against schistosomes in vitro and in vivo. Mefloquine–praziquantel and mefloquine–artesunate–praziquantel combinations were tested in clinical trials against chronic *Schistosoma haematobium* infection. Unfortunately, adverse event rates were high and there was no increased efficacy compared to praziquantel[106]. In the case of filarial worms, an alternative strategy is targeting of the *Wolbachia* endosymbiotic bacteria; this results in slow killing of the worms, thereby avoiding serious adverse events. For example, four weeks of treatment with doxycycline cured people suffering from onchocerciasis. The major aim of *Wolbachia*-directed therapy is to find shorter-course antibiotics[107]. Several high-throughput phenotypic screens have been carried out to discover new chemical entities that act against helminth infections[108,109], but no new clinical candidates have been reported to date.

Fungal infections

Fungal infections are particularly serious in immunocompromized patients. In addition to the common fungal infections, several are found predominantly in tropical regions. One of the most important problems is meningitis due to *C. neoformans*, which is estimated to cause 180,000 deaths per year among people infected with HIV, the majority of whom are in Africa[110]. This is poorly treated at present, and there is a need for new more efficacious drugs. Mycetoma, a disfiguring infection of the skin and soft tissue, is a particularly neglected disease, and is endemic to tropical and subtropical areas. Mycetoma can be caused by bacteria (actinomycetoma) or by fungi (eumycetoma). Fungal mycetoma requires long treatment with azoles, but drug resistance, disease recurrence and side effects are common. Patients often develop deformities that require amputations. DNDi are currently running a phase II/III trial for fosravuconazole to assess whether this azole is more effective and safer than the the currently used azoles itraconazole and ketoconazole. A comprehensive review of the current antifungal pipeline has been published recently[111].

Conclusion

The current treatments used against tropical diseases are suboptimal, and in some cases, there are no drugs available. However, recent progress in drug discovery in human African trypanosomiasis, tuberculosis and malaria show that headway can be made with the concerted efforts of governments, charities, foundations, product-development partnerships, academic institutions and pharmaceutical companies. There is still a long way to go, even in these disease areas. The high attrition rate in clinical trials and issues of resistance mean that there is no room for complacency. Further, multiple new agents are required for each disease area to allow combination therapies. While there has been progress in some disease areas, there is much less support and activity in others.

Received: 8 January 2018; Accepted: 2 May 2018;
Published online 25 July 2018.

1. World Health Organization. *Global Health Estimates 2015: Deaths by Cause, Age, Sex, by Country and by Region, 2000–2015*. (World Health Organization, Geneva, 2016).
2. World Health Organization. *Global Health Estimates 2015: Disease burden by Cause, Age, Sex, by Country and by Region, 2000–2015*. (World Health Organization, Geneva, 2016).
3. World Health Organization. *Neglected Tropical Diseases* http://www.who.int/neglected_diseases/diseases/en/ (2018).
4. Shepard, D. S., Undurraga, E. A., Halasa, Y. A. & Stanaway, J. D. The global economic burden of dengue: a systematic analysis. *Lancet Infect. Dis.* **16**, 935–941 (2016).
5. Conteh, L., Engels, T. & Molyneux, D. H. Socioeconomic aspects of neglected tropical diseases. *Lancet* **375**, 239–247 (2010).
6. G-Finder. *Neglected Disease Research and Development: Reflecting on a Decade of Global Investment*. (Policy Cures Research, Brisbane, 2017).
7. IPFMA. *The Pharmaceutical Industry and Global Health. Facts and Figures 2017*. (International Federation of Pharmaceutical Manufacturers & Associations, Geneva, 2017).
8. Wyatt, P. G., Gilbert, I. H., Read, K. D. & Fairlamb, A. H. Target validation: linking target and chemical properties to desired product profile. *Curr. Top. Med. Chem.* **11**, 1275–1283 (2011).
9. Burrows, J. N. *et al*. New developments in anti-malarial target candidate and product profiles. *Malar. J.* **16**, 26 (2017).
A comprehensive discussion of target product and target candidate profiles for malaria.
10. DNDi. Diseases & Projects. https://www.dndi.org/diseases-projects (2018).
11. Chatelain, E. Chagas disease drug discovery: toward a new era. *J. Biomol. Screen.* **20**, 22–35 (2015).
12. Huston, C. D. et al. A proposed target product profile and developmental cascade for new cryptosporidiosis treatments. *PLoS Negl. Trop. Dis.* **9**, e0003987 (2015).
13. Whitehorn, J. et al. Dengue therapeutics, chemoprophylaxis, and allied tools: state of the art and future directions. *PLoS Negl. Trop. Dis.* **8**, e3025 (2014).
14. Pawlowic, M. C., Vinayak, S., Sateriale, A., Brooks, C. F. & Striepen, B. Generating and maintaining transgenic *Cryptosporidium parvum* parasites. *Curr. Protoc. Microbiol.* **46**, 20B.2.1–20B.2.32 (2017).
15. Bushell, E. et al. Functional profiling of a *Plasmodium* genome reveals an abundance of essential genes. *Cell* **170**, 260–272 (2017).
16. de Koning-Ward, T. F., Gilson, P. R. & Crabb, B. S. Advances in molecular genetic systems in malaria. *Nat. Rev. Microbiol.* **13**, 373–387 (2015).
17. Lander, N., Chiurillo, M. A. & Docampo, R. Genome editing by CRISPR/Cas9: a game change in the genetic manipulation of protists. *J. Eukaryot. Microbiol.* **63**, 679–690 (2016).
18. Wang, Y. et al. A 'suicide' CRISPR–Cas9 system to promote gene deletion and restoration by electroporation in *Cryptococcus neoformans*. *Sci. Rep.* **6**, 31145 (2016).
19. Zheng, W., Thorne, N. & McKew, J. C. Phenotypic screens as a renewed approach for drug discovery. *Drug Discov. Today* **18**, 1067–1073 (2013).

© 2018 Springer Nature Limited. All rights reserved.

20. Ang, M. L. & Pethe, K. Contribution of high-content imaging technologies to the development of anti-infective drugs. *Cytometry A* **89**, 755–760 (2016).
21. Horvath, P. et al. Screening out irrelevant cell-based models of disease. *Nat. Rev. Drug Discov.* **15**, 751–769 (2016).
22. Fux, C. A., Shirtliff, M., Stoodley, P. & Costerton, J. W. Can laboratory reference strains mirror 'real-world' pathogenesis? *Trends Microbiol.* **13**, 58–63 (2005).
23. Angulo-Barturen, I. et al. A murine model of *falciparum*-malaria by in vivo selection of competent strains in non-myelodepleted mice engrafted with human erythrocytes. *PLoS ONE* **3**, e2252 (2008).
Description of a humanized mouse model for *P. falciparum* malaria.
24. Mikolajczak, S. A. et al. *Plasmodium vivax* liver stage development and hypnozoite persistence in human liver-chimeric mice. *Cell Host Microbe* **17**, 526–535 (2015).
25. Zhan, L. J., Tang, J., Sun, M. M. & Qin, C. Animal models for tuberculosis in translational and precision medicine. *Front. Microbiol.* **8**, 717 (2017).
26. Kramnik, I. & Beamer, G. Mouse models of human TB pathology: roles in the analysis of necrosis and the development of host-directed therapies. *Semin. Immunopathol.* **38**, 221–237 (2016).
27. Myllymaki, H., Niskanen, M., Oksanen, K. E. & Ramet, M. Animal models in tuberculosis research—where is the beef? *Expert Opin. Drug Discov.* **10**, 871–883 (2015).
28. Francisco, A. F. et al. Limited ability of posaconazole to cure both acute and chronic *Trypanosoma cruzi* infections revealed by highly sensitive in vivo imaging. *Antimicrob. Agents Chemother.* **59**, 4653–4661 (2015).
29. White, N. J. Determinants of relapse periodicity in *Plasmodium vivax* malaria. *Malar. J.* **10**, 297 (2011).
30. Barry, C. E. III et al. The spectrum of latent tuberculosis: rethinking the biology and intervention strategies. *Nat. Rev. Microbiol.* **7**, 845 (2009).
31. Sanchez-Valdez, F. J., Padilla, A., Wang, W., Orr, D. & Tarleton, R. L. Spontaneous dormancy protects *Trypanosoma cruzi* during extended drug exposure. *eLife* **7**, e34039 (2018).
32. Silver, L. L. A Gestalt approach to Gram-negative entry. *Bioorg. Med. Chem.* **24**, 6379–6389 (2016).
A description of the barriers to compound penetration in Gram-negative bacteria.
33. Brown, D. G., May-Dracka, T. L., Gagnon, M. M. & Tommasi, R. Trends and exceptions of physical properties on antibacterial activity for Gram-positive and Gram-negative pathogens. *J. Med. Chem.* **57**, 10144–10161 (2014).
34. Sarathy, J. P. et al. Prediction of drug penetration in tuberculosis lesions. *ACS Infect. Dis.* **2**, 552–563 (2016).
35. Fromm, M. F. Importance of P-glycoprotein at blood–tissue barriers. *Trends Pharmacol. Sci.* **25**, 423–429 (2004).
36. Woodford, N. & Ellington, M. J. The emergence of antibiotic resistance by mutation. *Clin. Microbiol. Infect.* **13**, 5–18 (2007).
37. Irwin, K. K., Renzette, N., Kowalik, T. F. & Jensen, J. D. Antiviral drug resistance as an adaptive process. *Virus Evol.* **2**, vew014 (2016).
38. Blasco, B., Leroy, D. & Fidock, D. A. Antimalarial drug resistance: linking *Plasmodium falciparum* parasite biology to the clinic. *Nat. Med.* **23**, 917–928 (2017).
39. Hammer, S. M. et al. Treatment for adult HIV infection: 2006 recommendations of the international AIDS society—USA panel. *J. Am. Med. Assoc.* **296**, 827–843 (2006).
40. World Health Organization. *World Malaria Report 2017*. (World Health Organization, Geneva, 2017).
41. Wells, T. N., Hooft van Huijsduijnen, R. & Van Voorhis, W. C. Malaria medicines: a glass half full? *Nat. Rev. Drug Discov.* **14**, 424–442 (2015).
42. White, N. J. et al. Antimalarial activity of KAF156 in *falciparum* and *vivax* malaria. *N. Engl. J. Med.* **375**, 1152–1160 (2016).
43. Macintyre, F. et al. A randomised, double-blind clinical phase II trial of the efficacy, safety, tolerability and pharmacokinetics of a single dose combination treatment with artefenomel and piperaquine in adults and children with uncomplicated *Plasmodium falciparum* malaria. *BMC Med.* **15**, 181 (2017).
44. White, N. J. et al. Spiroindolone KAE609 for *falciparum* and *vivax* malaria. *N. Engl. J. Med.* **371**, 403–410 (2014).
45. Paquet, T. et al. Antimalarial efficacy of MMV390048, an inhibitor of *Plasmodium* phosphatidylinositol 4-kinase. *Sci. Transl. Med.* **9**, eaad9735 (2017).
46. Lim, M. Y. et al. UDP-galactose and acetyl-CoA transporters as *Plasmodium* multidrug resistance genes. *Nat. Microbiol*, **1**, 16166 (2016).
47. Jimenez-Diaz, M. B. et al. (+)-SJ733, a clinical candidate for malaria that acts through ATP4 to induce rapid host-mediated clearance of *Plasmodium*. *Proc. Natl Acad. Sci. USA* **111**, E5455–E5462 (2014).
48. Baragana, B. et al. A novel multiple-stage antimalarial agent that inhibits protein synthesis. *Nature* **522**, 315–320 (2015).
49. McCarthy, J. S. et al. Safety, tolerability, pharmacokinetics, and activity of the novel long-acting antimalarial DSM265: a two-part first-in-human phase 1a/1b randomised study. *Lancet Infect. Dis.* **17**, 626–635 (2017).
A description of a clinical trial for an antimalarial that includes a human challenge trial.
50. Cowell, A. N. et al. Mapping the malaria parasite druggable genome by using in vitro evolution and chemogenomics. *Science* **359**, 191–199 (2018).
51. Plouffe, D. M. et al. High-throughput assay and discovery of small molecules that interrupt malaria transmission. *Cell Host Microbe* **19**, 114–126 (2016).
52. Hovlid, M. L. & Winzeler, E. A. Phenotypic screens in antimalarial drug discovery. *Trends Parasitol.* **32**, 697–707 (2016).
53. McCarthy, J. S. et al. Linking murine and human *Plasmodium falciparum* challenge models in a translational path for antimalarial drug development. *Antimicrob. Agents Chemother.* **60**, 3669–3675 (2016).
54. Stanisic, D. I., McCarthy, J. S. & Good, M. F. Controlled human malaria infection: applications, advances, and challenges. *Infect. Immun.* **86**, e00479-17 (2018).
55. Amadi, B. et al. High dose prolonged treatment with nitazoxanide is not effective for cryptosporidiosis in HIV positive Zambian children: a randomised controlled trial. *BMC Infect. Dis.* **9**, 195 (2009).
56. Manjunatha, U. H., Chao, A. T., Leong, F. J. & Diagana, T. T. Cryptosporidiosis drug discovery: opportunities and challenges. *ACS Infect. Dis.* **2**, 530–537 (2016).
57. Hulverson, M. A. et al. Bumped-kinase inhibitors for cryptosporidiosis therapy. *J. Infect. Dis.* **215**, 1275–1284 (2017).
Description of development of a series of compounds against cryptosporidiosis.
58. Manjunatha, U. H. et al. A *Cryptosporidium* PI(4)K inhibitor is a drug candidate for cryptosporidiosis. *Nature* **546**, 376–380 (2017).
Description of the discovery of a preclinical candidate for cryptosporidiosis by re-purposing of a series initially developed for malaria.
59. Love, M. S. et al. A high-throughput phenotypic screen identifies clofazimine as a potential treatment for cryptosporidiosis. *PLoS Negl. Trop. Dis.* **11**, e0005373 (2017).
60. Field, M. C. et al. Anti-trypanosomatid drug discovery: an ongoing challenge and a continuing need. *Nat. Rev. Microbiol.* **15**, 217–231 (2017).
A review encompassing drug discovery for the kinetoplastids.
61. Buscher, P., Cecchi, G., Jamonneau, V. & Priotto, G. Human African trypanosomiasis. *Lancet* **390**, 2397–2409 (2017).
62. Priotto, G. et al. Nifurtimox-eflornithine combination therapy for second-stage African *Trypanosoma brucei gambiense* trypanosomiasis: a multicentre, randomised, phase III, non-inferiority trial. *Lancet* **374**, 56–64 (2009).
63. Njamnshi, A. K., Gettinby, G. & Kennedy, P. G. E. The challenging problem of disease staging in human African trypanosomiasis (sleeping sickness): a new approach to a circular question. *Trans. R. Soc. Trop. Med. Hyg.* **111**, 199–203 (2017).
64. Steinmann, P., Stone, C. M., Sutherland, C. S., Tanner, M. & Tediosi, F. Contemporary and emerging strategies for eliminating human African trypanosomiasis due to *Trypanosoma brucei gambiense*. *Trop. Med. Int. Health* **20**, 707–718 (2015).
65. Torreele, E. et al. Fexinidazole—a new oral nitroimidazole drug candidate entering clinical development for the treatment of sleeping sickness. *PLoS Negl. Trop. Dis.* **4**, e923 (2010).
66. Mesu, V. et al. Oral fexinidazole for late-stage African *Trypanosoma brucei gambiense* trypanosomiasis: a pivotal multicentre, randomised, non-inferiority trial. *Lancet* **391**, 144–154 (2018).
Results from a human African trypanosomiasis clinical trial demonstrating non-inferiority of fexinidazole compared to NECT treatment.
67. Jacobs, R. T. et al. SCYX-7158, an orally-active benzoxaborole for the treatment of stage 2 human African trypanosomiasis. *PLoS Negl. Trop. Dis.* **5**, e1151 (2011).
68. Singh, O. P., Hasker, E., Boelaert, M. & Sundar, S. Elimination of visceral leishmaniasis on the Indian subcontinent. *Lancet Infect. Dis.* **16**, e304–e309 (2016).
69. DNDi. *DNDi Portfolio 2017* https://www.dndi.org/diseases-projects/portfolio (2017).
70. Khare, S. et al. Proteasome inhibition for treatment of leishmaniasis, Chagas disease and sleeping sickness. *Nature* **537**, 229–233 (2016).
71. Chatelain, E. Chagas disease research and development: is there light at the end of the tunnel? *Comput. Struct. Biotechnol. J.* **15**, 98–103 (2017).
72. Molina, I. et al. Randomized trial of posaconazole and benznidazole for chronic Chagas' disease. *N. Engl. J. Med.* **370**, 1899–1908 (2014).
73. Ford, N. et al. The WHO public health approach to HIV treatment and care: looking back and looking ahead. *Lancet Infect. Dis.* **18**, e76–e86 (2018).
74. Meanwell, N. A. 2015 Philip S. Portoghese Medicinal Chemistry Lectureship. Curing hepatitis C virus infection with direct-acting antiviral agents: the arc of a medicinal chemistry triumph. *J. Med. Chem.* **59**, 7311–7351 (2016).
75. Basile, K., Kok, J. & Dwyer, D. E. Zika virus: what, where from and where to? *Pathology* **49**, 698–706 (2017).
76. Wahid, B., Ali, A., Rafique, S. & Idrees, M. Global expansion of chikungunya virus: mapping the 64-year history. *Int. J. Infect. Dis.* **58**, 69–76 (2017).
77. Gyawali, N., Bradbury, R. S. & Taylor-Robinson, A. W. The epidemiology of dengue infection: harnessing past experience and current knowledge to support implementation of future control strategies. *J. Vector Borne Dis.* **53**, 293–304 (2016).
78. Coltart, C. E., Lindsey, B., Ghinai, I., Johnson, A. M. & Heymann, D. L. The Ebola outbreak, 2013–2016: old lessons for new epidemics. *Phil. Trans. R. Soc. Lond. B* **372**, 20160297 (2017).
79. de Wit, E., van Doremalen, N., Falzarano, D. & Munster, V. J. SARS and MERS: recent insights into emerging coronaviruses. *Nat. Rev. Microbiol.* **14**, 523–534 (2016).
80. Boldescu, V., Behnam, M. A. M., Vasilakis, N. & Klein, C. D. Broad-spectrum agents for flaviviral infections: dengue, Zika and beyond. *Nat. Rev. Drug Discov.* **16**, 565–586 (2017).
81. Sissoko, D. et al. Experimental treatment with favipiravir for Ebola virus disease (the JIKI Trial): a historically controlled, single-arm proof-of-concept trial in Guinea. *PLoS Med.* **13**, e1001967 (2016).
82. Warren, T. K. et al. Therapeutic efficacy of the small molecule GS-5734 against Ebola virus in rhesus monkeys. *Nature* **531**, 381–385 (2016).
83. Check Hayden, E. Experimental drugs poised for use in Ebola outbreak. *Nature* **557**, 475–476 (2018).

© 2018 Springer Nature Limited. All rights reserved.

84. Taylor, R. et al. BCX4430 — a broad-spectrum antiviral adenosine nucleoside analog under development for the treatment of Ebola virus disease. *J. Infect. Public Health* **9**, 220–226 (2016).
85. Warren, T. K. et al. Protection against filovirus diseases by a novel broad-spectrum nucleoside analogue BCX4430. *Nature* **508**, 402–405 (2014).
86. Debing, Y., Neyts, J. & Delang, L. The future of antivirals: broad-spectrum inhibitors. *Curr. Opin. Infect. Dis.* **28**, 596–602 (2015).
87. Hoagland, D. T., Liu, J., Lee, R. B. & Lee, R. E. New agents for the treatment of drug-resistant *Mycobacterium tuberculosis*. *Adv. Drug Deliv. Rev.* **102**, 55–72 (2016).
88. Lamprecht, D. A. et al. Turning the respiratory flexibility of *Mycobacterium tuberculosis* against itself. *Nat. Commun.* **7**, 12393 (2016).
89. Kakkar, A. K. & Dahiya, N. Bedaquiline for the treatment of resistant tuberculosis: promises and pitfalls. *Tuberculosis* **94**, 357–362 (2014).
90. Matsumoto, M. et al. OPC-67683, a nitro-dihydro-imidazooxazole derivative with promising action against tuberculosis in vitro and in mice. *PLoS Med.* **3**, e466 (2006).
91. Olaru, I. D. et al. Novel drugs against tuberculosis: a clinician's perspective. *Eur. Respir. J.* **45**, 1119–1131 (2015).
92. Lee, M. et al. Linezolid for XDR-TB—final study outcomes. *N. Engl. J. Med.* **373**, 290–291 (2015).
93. Li, X. et al. Discovery of a potent and specific *M. tuberculosis* leucyl-tRNA synthetase inhibitor: (*S*)-3-(aminomethyl)-4-chloro-7-(2-hydroxyethoxy) benzo[c][1,2]oxaborol-1(3*H*)-ol (GSK656). *J. Med. Chem.* **60**, 8011–8026 (2017).
94. World Health Organization. *Global Tuberculosis Report 2017*. (World Health Organization, Geneva, 2017).
95. Tacconelli, E. et al. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infect. Dis.* **18**, 318–327 (2018).
96. World Health Organization. *Antibacterial Agents in Clinical Development: an Analysis of the Antibacterial Clinical Development Pipeline, Including Tuberculosis*. (World Health Organization, Geneva, 2017).
97. Baker, S. et al. Fitness benefits in fluoroquinolone-resistant *Salmonella* Typhi in the absence of antimicrobial pressure. *eLife* **2**, e01229 (2013).
98. Wong, V. K. et al. Phylogeographical analysis of the dominant multidrug-resistant H58 clade of *Salmonella* Typhi identifies inter- and intracontinental transmission events. *Nat. Genet.* **47**, 632–639 (2015).
99. Hotez, P. J. et al. Helminth infections: the great neglected tropical diseases. *J. Clin. Invest.* **118**, 1311–1321 (2018).
100. Moser, W. et al. Efficacy and safety of tribendimidine, tribendimidine plus ivermectin, tribendimidine plus oxantel pamoate, and albendazole plus oxantel pamoate against hookworm and concomitant soil-transmitted helminth infections in Tanzania and Cote d'Ivoire: a randomised, controlled, single-blinded, non-inferiority trial. *Lancet Infect. Dis.* **17**, 1162–1171 (2017).
101. Sayasone, S. et al. Efficacy and safety of tribendimidine versus praziquantel against *Opisthorchis viverrini* in Laos: an open-label, randomised, non-inferiority, phase 2 trial. *Lancet Infect. Dis.* **18**, 155–161 (2018).
102. Xiao, S. H., Utzinger, J., Tanner, M., Keiser, J. & Xue, J. Advances with the Chinese anthelminthic drug tribendimidine in clinical trials and laboratory investigations. *Acta Trop.* **126**, 115–126 (2013).
A comprehensive review of the discovery and development of tribendimidine, the first new anthelminthic in 30 years.
103. Moser, W. et al. Efficacy and safety of oxantel pamoate in school-aged children infected with *Trichuris trichiura* on Pemba Island, Tanzania: a parallel, randomised, controlled, dose-ranging study. *Lancet Infect. Dis.* **16**, 53–60 (2016).
104. Moser, W. et al. Efficacy and tolerability of triple drug therapy with albendazole, pyrantel pamoate, and oxantel pamoate compared with albendazole plus oxantel pamoate, pyrantel pamoate plus oxantel pamoate, and mebendazole plus pyrantel pamoate and oxantel pamoate against hookworm infections in school-aged children in Laos: a randomised, single-blind trial. *Lancet Infect. Dis.* **18**, 729–737 (2018).
105. Awadzi, K., Opoku, N. O., Attah, S. K., Lazdins-Helds, J. & Kuesel, A. C. A randomized, single-ascending-dose, ivermectin-controlled, double-blind study of moxidectin in *Onchocerca volvulus* infection. *PLoS Negl. Trop. Dis.* **8**, e2953 (2014).
106. Keiser, J. et al. Praziquantel, mefloquine–praziquantel, and mefloquine–artesunate–praziquantel against *Schistosoma haematobium*: a randomized, exploratory, open-label trial. *PLoS Negl. Trop. Dis.* **8**, e2975 (2014).
107. Klarmann-Schulz, U. et al. Comparison of doxycycline, minocycline, doxycycline plus albendazole and albendazole alone in their efficacy against onchocerciasis in a randomized, open-label, pilot trial. *PLoS Negl. Trop. Dis.* **11**, e0005156 (2017).
108. Guidi, A. et al. Discovery by organism based high-throughput screening of new multi-stage compounds affecting *Schistosoma mansoni* viability, egg formation and production. *PLoS Negl. Trop. Dis.* **11**, e0005994 (2017).
109. Mansour, N. R. et al. High throughput screening identifies novel lead compounds with activity against larval, juvenile and adult *Schistosoma mansoni*. *PLoS Negl. Trop. Dis.* **10**, e0004659 (2016).
110. Rajasingham, R. et al. Global burden of disease of HIV-associated cryptococcal meningitis: an updated analysis. *Lancet Infect. Dis.* **17**, 873–881 (2017).
111. Perfect, J. R. The antifungal pipeline: a reality check. *Nat. Rev. Drug Discov.* **16**, 603–616 (2017).
112. Kotloff, K. L. et al. Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case–control study. *Lancet* **382**, 209–222 (2013).
The results of a clinical study into enteric diseases in LMICs.
113. Liu, J. et al. Use of quantitative molecular diagnostic methods to identify causes of diarrhoea in children: a reanalysis of the GEMS case–control study. *Lancet* **388**, 1291–1301 (2016).
114. Hotez, P. J., Fenwick, A., Savioli, L. & Molyneux, D. H. Rescuing the bottom billion through control of neglected tropical diseases. *Lancet* **373**, 1570–1575 (2009).
115. Checkley, W. et al. A review of the global burden, novel diagnostics, therapeutics, and vaccine targets for cryptosporidium. *Lancet Infect. Dis.* **15**, 85–94 (2015).
116. Kublin, J. G. et al. Effect of *Plasmodium falciparum* malaria on concentration of HIV-1-RNA in the blood of adults in rural Malawi: a prospective cohort study. *Lancet* **365**, 233–240 (2005).
117. Mayer, K. H. & Dukes Hamilton, C. Synergistic pandemics: confronting the global HIV and tuberculosis epidemics. *Clin. Infect. Dis.* **50**, S67–S70 (2010).
118. Brooker, S. et al. Epidemiology of *Plasmodium*–helminth co-infection in Africa: populations at risk, potential impact on anemia, and prospects for combining control. *Am. J. Trop. Med. Hyg.* **77**, 88–98 (2007).
119. Blanc, F. X. et al. Earlier versus later start of antiretroviral therapy in HIV-infected adults with tuberculosis. *N. Engl. J. Med.* **365**, 1471–1481 (2011).
120. Josephson, F. Drug–drug interactions in the treatment of HIV infection: focus on pharmacokinetic enhancement through CYP3A inhibition. *J. Intern. Med.* **268**, 530–539 (2010).
121. Pion, S. D. et al. Co-infection with *Onchocerca volvulus* and *Loa loa* microfilariae in central Cameroon: are these two species interacting? *Parasitology* **132**, 843–854 (2006).
122. Hoerauf, A., Pfarr, K., Mand, S., Debrah, A. Y. & Specht, S. Filariasis in Africa—treatment challenges and prospects. *Clin. Microbiol. Infect.* **17**, 977–985 (2011).
123. Song, T. et al. Linezolid trough concentrations correlate with mitochondrial toxicity-related adverse events in the treatment of chronic extensively drug-resistant tuberculosis. *EBioMedicine* **2**, 1627–1633 (2015).
124. Tiberi, S. et al. Tuberculosis: progress and advances in development of new drugs, treatment regimens, and host-directed therapies. *Lancet Infect. Dis* **18**, e183–e198 (2018).

Acknowledgements We acknowledge the following organizations that have supported our drug-discovery efforts for some of these diseases: Wellcome Trust, Medicines for Malaria Venture, the Bill and Melinda Gates Foundation and Drugs for Neglected Diseases initiative. We thank members of the Drug Discovery Unit for helpful discussions and K. Read, S. Green and S. Wyllie for reading the manuscript.

Reviewer information *Nature* thanks B. Melillo, E. Winzeler and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions All the authors contributed to writing the text.

Competing interests The authors declare no competing interests.

Additional information
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to I.H.G.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

REVIEW

https://doi.org/10.1038/s41586-018-0306-9

Acting on non-communicable diseases in low- and middle-income tropical countries

Majid Ezzati[1,2,3]*, Jonathan Pearson-Stuttard[1,2], James E. Bennett[1,2] & Colin D. Mathers[4]

The classical portrayal of poor health in tropical countries is one of infections and parasites, contrasting with wealthy Western countries, where unhealthy diet and behaviours cause non-communicable diseases (NCDs) such as heart disease and cancer. Using international mortality data, we show that most NCDs cause more deaths at every age in low- and middle-income tropical countries than in high-income Western countries. Causes of NCDs in low- and middle-income countries include poor nutrition and living environment, infections, insufficient taxation and regulation of tobacco and alcohol, and under-resourced and inaccessible healthcare. We identify a comprehensive set of actions across health, social, economic and environmental sectors that could confront NCDs in low- and middle-income tropical countries and reduce global health inequalities.

Poor health in the tropics often brings to mind infectious and parasitic diseases that thrive and spread in hot and humid climates (http://www.who.int/topics/tropical_diseases/en/ and http://www.astmh.org/education-resources/tropical-medicine-q-a/major-tropical-diseases), rather than NCDs like cancers and heart disease, which are considered ailments of wealthy urbanized Western nations[1,2]. NCDs refer to conditions other than injuries, maternal and perinatal conditions and nutritional deficiencies that do not involve an acute infection or parasite, and hence are not transmittable from person to person.

In the 1970s, Omran proposed[3] a theory of epidemiological transition in which, as countries develop economically and 'modernise', there is a shift from communicable (infectious) diseases to NCDs. This linear perspective on health development has led the bulk of domestic and international health financing and policy in low- and middle-income countries to focus on infectious diseases and maternal and child health[4].

As worldwide data on deaths and disease have become more available, it has become increasingly apparent that NCDs also affect the populations of low- and middle-income nations[5–7]. The growing evidence for the relevance of NCDs as a global health issue led to a United Nations high-level meeting on the prevention and control of NCDs, following which the World Health Organization (WHO) member states committed to reducing premature mortality from four NCDs (cancers, cardiovascular and chronic respiratory diseases, and diabetes in people in the age range 30–69 years) by a quarter relative to their 2010 levels by 2025. NCDs have also been included in the United Nations Sustainable Development Goals (SDGs), with a target of reducing by one-third, relative to their 2015 levels, premature mortality from NCDs and promoting mental health and well-being by 2030; the indicator used to measure progress towards this target is the same as that of the WHO target.

Despite increasing awareness of the importance of NCDs as a global health issue, there is limited evaluation of how similarities and differences in NCD composition between low- and middle-income tropical and high-income countries should influence policies and programmes that aim to reduce their health burden[5,8,9]. In this Review, we use international data to compare NCDs between low- and middle-income tropical and high-income Western regions (Box 1), and use these comparisons to identify policies and interventions that could reduce their burden. We begin by describing the overall mortality burden and composition of NCDs in tropical regions in comparison to high-income countries. We then evaluate some of the potential causes of regional differences of NCDs, including genetics, infections, nutrition, healthcare and poverty. Finally, we identify a comprehensive set of actions across health, social, economic and environmental sectors aimed at reducing NCDs in low- and middle-income countries.

NCD deaths in the tropics

In 2016, life expectancy at birth in low- and middle-income tropical countries ranged from less than 53 years in Sierra Leone and Central African Republic to above 79 years in some countries in Latin America and the Caribbean, for example Costa Rica and Cuba (Supplementary Fig. 1). With the exception of the USA, in which life expectancy at birth of 78.6 years was similar to that of some better-off low- and middle-income nations, life expectancy was consistently high in high-income Western countries, surpassing 83 years in Switzerland, Spain, France, Canada and Australia.

Of the 25.3 million deaths in the four tropical regions—Africa, India (south Asia), southeast Asia, and Latin America and the Caribbean—in 2016, 8.7 million (34%) were from infectious and parasitic diseases, maternal and perinatal conditions, and nutritional deficiencies, 14.0 million (55%) from NCDs and the remainder from injuries. The share of deaths from NCDs in tropical countries had the same geographical pattern as life expectancy, ranging from below 30% in some African countries to over 80% in parts of southeast Asia and Latin America and the Caribbean, Mauritius and Seychelles (Fig. 1a). Tropical Africa was the only region in which NCDs accounted for fewer deaths than infectious and parasitic diseases, maternal and perinatal conditions, and nutritional deficiencies (33% versus 57%). NCDs accounted for nearly 90% of all deaths in high-income Western countries, higher than their share in all tropical regions, as predicted by the epidemiological transition theory. Crude death rates from NCDs as a whole, and from

[1]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK. [2]MRC-PHE Centre for Environment and Health, Imperial College London, London, UK. [3]WHO Collaborating Centre on NCD Surveillance and Epidemiology, Imperial College London, London, UK. [4]Department of Information, Evidence and Research, World Health Organization, Geneva, Switzerland. *e-mail: majid.ezzati@imperial.ac.uk

© 2018 Springer Nature Limited. All rights reserved.

Box 1

Low- and middle-income tropical countries

Physicians and medical researchers have had a longstanding interest in health in the tropics because its climate and ecology can influence human health directly or by changing the viability and dynamics of disease vectors. In the tropics and elsewhere, health is also strongly influenced by the economic resources available to countries and individuals. In this Review we evaluate the status of NCDs by focusing on low- and middle-income tropical countries as a way to bridge two parallel bodies of work in global health: tropical diseases and NCDs.

Countries were classified as 'tropical' if half or more of their land area was between the Tropic of Cancer and Tropic of Capricorn. This classification excludes some low-income countries, such as Bangladesh and Pakistan. Similarly, the focus on low- and middle-income excludes some tropical countries, such as Singapore. For reporting, we organised low- and middle-income tropical countries into the following regions.

(1) Tropical Africa: Angola, Benin, Botswana, Burkina Faso, Burundi, Cabo Verde, Cameroon, Central African Republic, Chad, Comoros, Congo, Cote d'Ivoire, Djibouti, DR Congo, Equatorial Guinea, Eritrea, Ethiopia, Gabon, Gambia, Ghana, Guinea, Guinea Bissau, Kenya, Liberia, Madagascar, Malawi, Mali, Mauritania, Mauritius*, Mozambique, Namibia, Niger, Nigeria, Rwanda, Sao Tome and Principe*, Senegal, Seychelles*, Sierra Leone, Somalia, South Sudan, Sudan, Tanzania, Togo, Uganda, Zambia and Zimbabwe.

(2) Tropical south Asia: India

(3) Tropical southeast Asia: Brunei Darussalam, Cambodia, Indonesia, Lao PDR, Malaysia, Maldives, Myanmar, Philippines, Sri Lanka, Thailand, Timor-Leste and Vietnam.

(4) Tropical Latin America and the Caribbean: Antigua and Barbuda, Barbados, Belize, Bolivia, Brazil, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, El Salvador, Grenada, Guatemala, Guyana, Haiti, Honduras, Jamaica, Nicaragua, Panama, Paraguay*, Peru, Saint Lucia*, Saint Vincent and the Grenadines*, Suriname, Trinidad and Tobago and Venezuela

Countries and territories marked with an asterisk had data on mortality but not on cancer incidence. To simplify reporting, we also did not analyse a few other low- and middle-income tropical countries that were not in these four regions. These include Fiji, Papua New Guinea, Micronesia, Kiribati, Tonga, Samoa, Solomon Islands, Vanuatu, Yemen and Oman.

Throughout the paper, we compare health in the four low- and middle-income tropical regions with high-income Western countries, which include Australia, Austria, Belgium, Canada, Cyprus, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Israel, Italy, Luxembourg, Malta, Netherlands, New Zealand, Norway, Portugal, Spain, Sweden, Switzerland, UK and USA.

most individual NCDs, were also higher in high-income Western nations than in low- and middle-income tropical regions.

Crude death rates do not take into account that the populations of low- and middle-income countries are younger than in high-income countries (Supplementary Fig. 2). For example, 83% of tropical Africa's population and 70% of India's population is younger than 40 years, compared to 49% in high-income Western nations. Many infections disproportionately affect infants and children (such as diarrhoea, malaria and measles) or young and middle-aged adults (such as HIV and tuberculosis). In contrast, deaths from most NCDs increase with age (although there are exceptions such as congenital anomalies and haemoglobinopathies like sickle cell). This combination of distinct age patterns of deaths from NCDs versus infections and different age structures is the main reason for higher crude NCD death rates in high-income countries. In other words, low- and middle-income tropical countries have lower NCD death rates than high-income nations at least partly because their population is younger.

One way to compare the epidemiological (versus demographic) features of NCDs across countries is to fix the age structure to a reference (standard) population, a process known as age-standardization (this is done by taking a weighted average of age-specific death rates, with the weight for each age group being its share of standard population)[10]. Once this is done, NCDs as a whole, and most individual NCDs, have higher death rates in tropical countries than they do in high-income Western countries (Figs. 1b and 2). The exceptions to this pattern are cancers, mental, neurological and substance-abuse disorders, and endocrine, blood and immune disorders, for which age-standardized death rates are slightly lower or about the same in the tropical regions (Fig. 2). In other words, low- and middle-income tropical countries have a disadvantage in NCDs compared to their high-income Western counterparts, as they do for infectious diseases. The largest absolute disadvantage of tropical regions is that of cardiovascular diseases, followed by chronic respiratory diseases, diabetes, kidney and other genitourinary diseases, and liver cirrhosis and other digestive diseases. In the following sections, we examine the regional patterns of cardiovascular diseases and cancers, which are the largest two groups of NCDs.

That NCDs have a substantial mortality burden in low- and middle-income countries is consistent with the historical experience of high-income countries[11–13]. For example, in 1900, the ten leading causes of death in the USA included infectious diseases such as respiratory infections, influenza, tuberculosis and diarrhoea as well as heart and kidney diseases, stroke and cancers; by 1930, these latter diseases made up four of the five leading causes of death[11,13]. As life expectancy increased in high-income countries, age-standardized death rates declined for some NCDs (stroke and possibly kidney disease) but increased for others (cancers)[11,12].

Regional inequalities in cardiovascular diseases

Cardiovascular diseases accounted for 5.8 million deaths (23% of all deaths) in the four low- and middle-income tropical regions, with their share ranging from 13% in Africa to 32% in southeast Asia, compared to 2.2 million deaths (30% of all deaths) in high-income Western countries. Age-standardized death rates from cardiovascular diseases as a whole, as well as from most individual cardiovascular diseases, were higher in low- and middle-income tropical regions than in the high-income Western region (Fig. 3).

In relative terms, the largest regional mortality disparity was that of rheumatic heart disease, with death rates in the African and Asian tropical regions many times those in the high-income Western region. Rheumatic heart disease, which is caused by rheumatic fever from group A streptococcal disease, is a classic NCD of poor communities[9,14]. In 2016, it accounted for nearly 300,000 worldwide deaths, nearly half of which occurred in tropical Africa, India and southeast Asia. However, even in these regions, rheumatic heart disease accounted for no more than 4% of cardiovascular disease deaths in 2016.

Rheumatic heart disease was present at the beginning of the twentieth century in high-income countries, declining sharply and virtually disappearing by the 1970s, except in disadvantaged indigenous communities[14–16]. Much of this decline occurred before the introduction of antibiotics, owing to improvements in housing, hygiene, nutrition and primary healthcare[11,14,15]. More recent declines in middle-income tropical countries such as Costa Rica and Cuba have benefited from a combination of environmental interventions and antibiotic therapy[14]. Incidence remains high in sub-Saharan Africa and south Asia, where systematic prevention and treatment programmes are rare[9,17]. Other cardiovascular diseases associated with infections that are more common in low-income than in high-income countries include endomyocardial fibrosis, pericarditis and cardiomyopathy[9,18,19]. Like rheumatic heart disease, they accounted for a small share of cardiovascular disease deaths in 2016.

© 2018 Springer Nature Limited. All rights reserved.

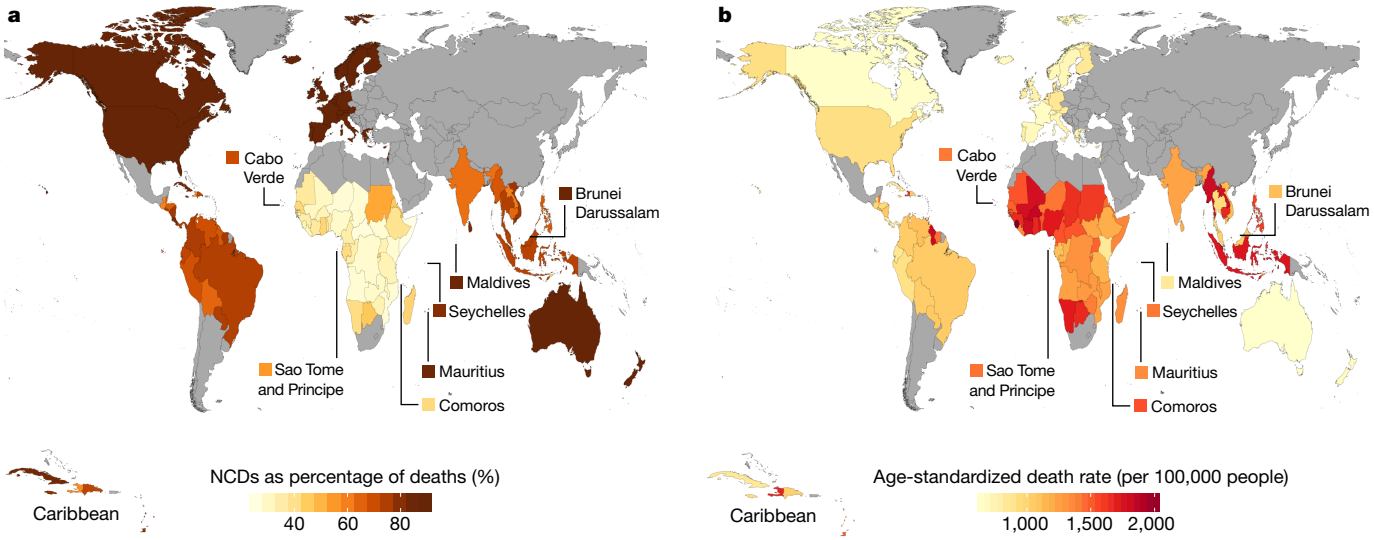


Fig. 1 | NCD mortality in low- and middle-income tropical countries and high-income Western countries. a, **b**, Share of deaths from NCDs (**a**) and age-standardized death rates from NCDs (**b**) in low- and middle-income tropical and high-income Western countries in 2016. NCDs include: cancers; cardiovascular diseases; diabetes; endocrine, blood and immune disorders; non-infectious respiratory, digestive (including liver), and genitourinary diseases; neurological conditions; mental and substance-use disorders; congenital anomalies; and sense-organ, skin, musculoskeletal and oral/dental conditions. The mortality and population data were taken from the WHO global health estimates[85]. All results are for females and males combined. Death rates were age-standardized to the WHO standard population[10].

Although the near-disappearance of rheumatic heart disease from high-income nations has, appropriately, put a spotlight on its persistent burden in the poorest countries, the largest absolute mortality inequality between tropical and high-income regions is unequivocally that of stroke and ischaemic heart disease (Fig. 3) because they have a large mortality burden—more than three-quarters of cardiovascular disease deaths in every region. Historical data from high-income countries consistently show a decline in stroke death rates from as early as 1900, driven initially by a sharp decline in haemorrhagic stroke and subsequently by both haemorrhagic and ischaemic subtypes[16]. Limited data from middle-income countries also show similar trends[16]. High blood pressure is the single most important risk factor for stroke, especially for haemorrhagic stroke[20], and is addressed later in this Review. Some stroke risk factors, such as sickle cell, are nonetheless predominantly tropical[21].

How ischaemic heart disease changes during economic and health development is more complex than trends for rheumatic heart disease and stroke, which both tend to decline as overall health improves[11,12,16,22]. In high-income countries, ischaemic heart disease mortality increased in the first half of the twentieth century and declined subsequently, with similar trends seen today in middle-income countries with longitudinal data including Seychelles, South Africa and parts of Latin America[11,16,22] Some studies have found a low prevalence of markers of coronary atherosclerosis, a precursor of ischaemic heart disease, in remote indigenous populations of the tropics—such as the Masai and Samburu in east Africa in the 1960s[23,24], the Kalahari Bushmen in the 1970s[25] and the Tsimane in South America in the 2010s[26]—even though diets, including the share of carbohydrates versus animal-source foods, vary substantially among these groups. It is, however, unclear how representative these groups are of the populations of the larger tropical regions where they live, and studies in the same and other tropical regions have found high levels of coronary disease or its precursors[24,27]. The variations in ischaemic heart disease rates and trends even within low- and middle-income countries may be because the disease has diverse causes, including chronic inflammation, fetal and early life malnutrition and stunted growth, smoking, heavy episodic (binge) alcohol drinking, and low consumption of fruits and vegetables, which are present to varying degrees in low- and middle-income countries. The large burden of ischaemic heart disease and stroke in low- and middle-income countries, and the similarities and differences in their causes compared to high-income countries, necessitate going beyond the simple paradigm of epidemiological transition to identify interventions that effectively address their local causes and health systems that are able to help patients with these conditions.

Diverse regional patterns of cancer mortality

Cancers accounted for 2.4 million deaths in 2016 (9% of all deaths) in the four low- and middle-income tropical regions, with their share ranging from 6% in Africa to 18% in Latin America and the Caribbean, compared to 1.9 million deaths (26% of all deaths) in high-income Western countries. Cancers are a heterogeneous group of diseases, so the regional variations may differ by type of cancer[8,28]. After age-standardization, death rates from many cancers were lower in the tropical regions than in high-income Western countries, with the difference largest for lung cancer, followed by colorectal and pancreatic cancers (Fig. 4). However, cervical, head and neck, liver, stomach and oesophagus cancers had higher death rates in all or some tropical regions than in the high-income Western region. Prostate cancer death rates were higher in tropical Africa and Latin America and the Caribbean than in high-income countries, but the opposite was true in India and southeast Asia.

The regional differences in cancer death rates are due to a combination of differences in incidence and survival. Variations in cancer incidence are, in turn, due to different exposures to its major risk factors including smoking and alcohol use, chemicals and radiation in the environment, infections, diet, reproductive and sexual practices, and metabolic and hormonal factors[28–31]. Access to and utilization of screening and treatment also matters when precancerous lesions or early-stage cancer can be effectively treated, such as for cervical cancers. The incidence of most cancers in low- and middle-income tropical regions was similar to, or lower than, that in high-income countries, with the difference largest for prostate and breast cancers, followed by lung and colorectal cancers (Supplementary Fig. 3). That prostate and breast cancers had lower incidence but higher death rates in low- and middle-income regions is because of worse cancer survival[32]. Cervical, liver and stomach cancers were the only cancers with higher age-standardized incidence in all or most low- and middle-income regions. Kaposi's sarcoma, of which HIV/AIDS is the leading cause, was unfamiliar to many physicians in the early 1980s, and has emerged as a leading cancer in Africa as a result of the HIV/AIDS epidemic, and the single most common cancer in countries such as Zimbabwe and Mozambique.

Almost all cervical cancers are caused by infection with human papillomavirus, which is mostly transmitted via sexual contact, and

© 2018 Springer Nature Limited. All rights reserved.

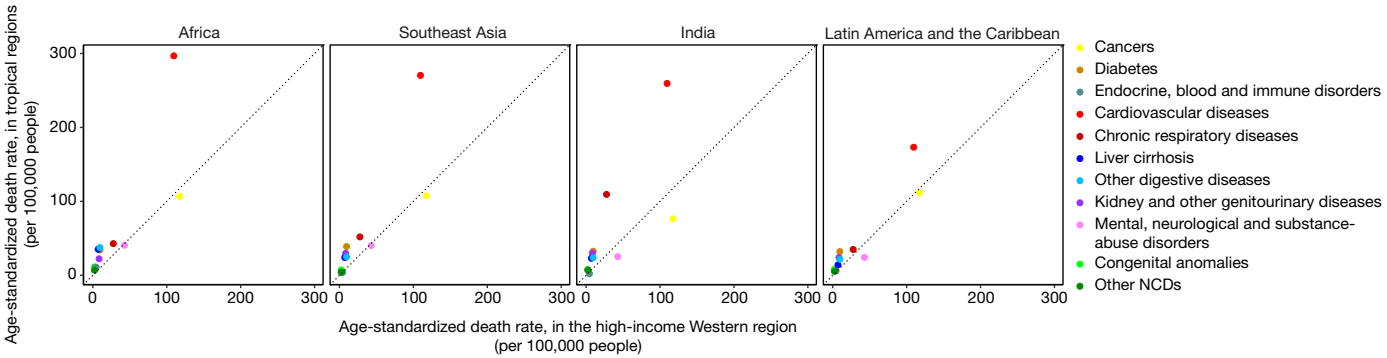


Fig. 2 | Comparison between low- and middle-income tropical regions and the high-income Western region of age-standardized death rates from NCDs in 2016. The mortality and population data were taken from the WHO global health estimates[85]. Results are for females and males combined. Death rates were age-standardized to the WHO standard population[10].

around 90% of non-cardia stomach cancers (that is, stomach cancer not including the upper stomach) are due to *Helicobacter pylori* infection (an estimated 70% of all stomach cancers are non-cardia)[33,34], a bacterium transmitted via contaminated food and water. Given their main routes of transmission, both infections are strongly related to the hygiene environment and behaviour. Cervical and stomach cancer incidence and deaths have declined steadily in high-income countries and in emerging economies, owing to improvements in water and sanitation, housing, electrification and refrigeration, and access to soap and other basic hygiene necessities, as well as effective screening and treatment programmes for cervical cancer[35–39]. Poorer countries fare worse in terms of the social, environmental and health-system determinants of these two infection-related cancers. Further, although an effective vaccine for human papillomavirus now exists, most lower-income nations have not introduced human papillomavirus vaccination, partly owing to its cost[28,40]. As a result, in 2016, nearly two thirds (63%) of worldwide deaths from cervical cancer deaths occurred in the four tropical regions covered in this Review, compared with 7% in high-income Western countries.

Liver cancer, which is also more common in low- and middle-income tropical countries than in high-income countries, has multiple causes, including hepatitis B and C viral infection, mycotoxins, especially aflatoxin from contamination of foods such as maize and groundnuts during storage, liver flukes, which are transmitted by eating infected raw freshwater fish, smoking, alcohol use, adiposity and diabetes[28,41–43]. In high-income countries, hepatitis B virus infection has declined, while adiposity and diabetes have increased[43–45], and alcohol use and smoking have declined in some countries and increased in others[46]. Owing to these complex trends, liver cancer declined in the middle decades of the twentieth century in high-income countries, followed by a continued decline in some countries and an increase in others, such as the USA and the UK[36]. Currently, the highest worldwide hepatitis B infection rates are those in tropical Africa, where hepatitis B virus vaccine coverage, especially for the birth dose, is lowest[42]. As a result, despite the availability of an effective vaccine, hepatitis B infection, in combination with aflatoxin, remains a key determinant of the large burden of liver cancer in Africa[28], especially in west Africa.

Although infections remain an important cause of cancers in low- and middle-income tropical countries, other cancer risk factors are increasingly important. In particular, smoking accounts for about 20% of all cancer deaths in low- and middle-income countries, compared to 30% in high-income countries[29]. However, smoking is rapidly shifting from high-income to low- and middle-income countries owing to differences in taxes and other tobacco control measures, and fuelled by the tobacco industry's tactics[47]. Harmful alcohol use, which accounts for about 5% of all cancer deaths worldwide is also common in Latin America, and rising in Africa and Asia[48]. The coexistence of cancers that are largely or partly caused by infections and those that are related to smoking and alcohol use means that reducing cancer burden in low- and middle-income countries must include social and environmental improvements such as clean water and sanitation, electrification and refrigeration to reduce the risk of infections, vaccination to protect against infections, and effective tobacco and alcohol control. However, the benefits of some of these actions will take decades to materialize. Therefore, health systems should be equipped to detect and treat precancerous signs or early-stage cancers[49].

Causes of NCDs in low- and middle-income countries

The data presented in this Review demonstrate that not only are NCDs present in low- and middle-income tropical countries, but that also, after accounting for differences in age structure, the populations of

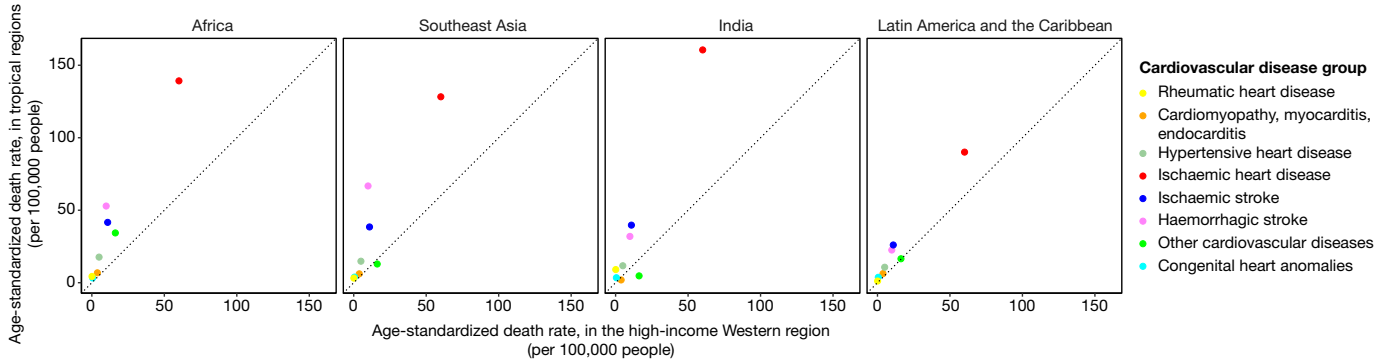


Fig. 3 | Age-standardized cardiovascular disease death rates in low- and middle-income tropical regions and the high-income Western region in 2016. A point above the dotted line has a higher death rate in the tropical region than in the high-income Western region, and a point below the dotted line has a lower death rate. Data are for females and males combined. Congenital heart anomalies are not classified as a cardiovascular disease by the international classification of disease system. They are shown here alongside the international classification of disease cardiovascular disease categories to show regional variations in their death rates[9]. Mortality and population data are taken from the WHO global health estimates[85]. Death rates were age-standardized to the WHO standard population[10].

© 2018 Springer Nature Limited. All rights reserved.

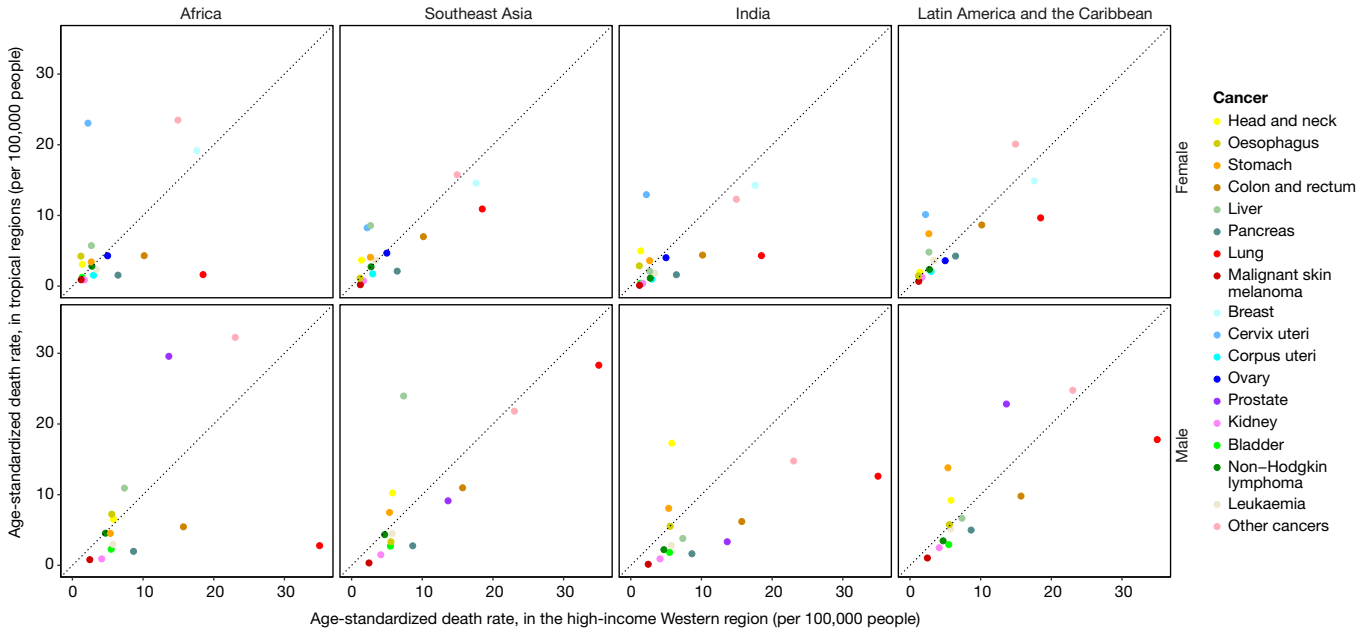


Fig. 4 | Age-standardized site-specific cancer death rates in tropical regions and the high-income Western region in 2016. A point above the dotted line has a higher death rate in the tropical region than in the high-income Western region, and a point below the dotted line has a lower death rate. Mortality and population data were taken from the WHO global health estimates[85] Death rates were age-standardized to the WHO standard population[10]. Kaposi's sarcoma is not shown as a separate category because its deaths were reported together with those from HIV/AIDS in the WHO global health estimates. See Supplementary Fig. 3 for incidence.

these countries experience a higher burden of NCDs than high-income Western countries. Beyond a general disadvantage, there are differences in the composition of NCDs across low- and middle-income tropical regions compared to the high-income Western region. Here we examine some of the potential determinants of these regional differences.

Genetics

Genetic variants that increase the risk of NCDs may have different regional frequencies. For example, the sickle haemoglobin gene (*HbS*), which is protective against malaria, is more frequent in tropical regions, especially in Africa, than in other parts of the world, with three-quarters of children with a haemoglobinopathy born in sub-Saharan Africa[50]. Within Africa, the frequency of the *HbS* gene is associated with the historical geographical distribution of malaria, supporting the hypothesis that a classical tropical disease has shaped the geographical distribution of an NCD via genetic selection[51,52].

It is possible that genetic susceptibility loci for other NCDs are also over- or under-represented in tropical regions. For example a variant on chromosome 17q21, which is more common in men of African descent than in non-African men, is associated with an increased risk of prostate cancer[53,54]. How much this single-nucleotide polymorphism accounts for the higher rates of prostate cancer in tropical Africa, however, is unclear because its causal effect is not fully established and its frequency varies across Africa. More broadly, currently known genetic variants do not account for the large regional variations in NCD rates seen earlier or for the fact that disease rates change over periods of years and decades. Nor do genetic variants explain why migrants from the tropics to Western countries tend to converge to disease rates in the host environment within a few generations. Therefore, it is important to evaluate genetics alongside social and environmental determinants in studying worldwide disease patterns.

Obesity and other cardiometabolic risks

One of the factors underlying the claims that NCDs, especially ischaemic heart disease and stroke, are uncommon in lower-income tropical countries is the assumption that diet worsens as societies become more affluent and urbanized, leading to higher rates of obesity and diabetes, and higher blood pressure and lipids[1,7,55–58]. However, affluence and urban living can lead to improvements as well as worsening in diet. For example, consumption of fruits, which are beneficial for NCDs, is seasonal in poorer countries, while more affluent populations use food trade, transportation and storage to consume fruits year-round and hence have higher average consumption[16,46,59].

Body-mass index (BMI) and associated metabolic conditions also have varied patterns in relation to national income[60]. BMI is currently lower in tropical Africa and south and southeast Asia than in the high-income Western region, but is about the same in the latter region and Latin America and the Caribbean (Fig. 5)[45]. In contrast to BMI, diabetes prevalence is about the same, or higher, in most low- and middle-income regions than in the high-income Western region[44]. The regional discordance between diabetes and elevated BMI, which is an important cause of diabetes, might be partly due to phenotypic variations arising from inadequate fetal and childhood nutrition and growth in low- and middle-income countries as discussed below, leading to earlier onset of β-cell dysfunction. Additionally, the better-resourced health systems in high-income countries can identify people at high risk of diabetes at an earlier stage, and use lifestyle and dietary modification and drugs to delay or prevent its onset.

In an even more extreme form of disadvantage in low- and middle-income regions, the highest worldwide blood pressure levels are those in sub-Saharan Africa, with blood pressure also high in south Asia, whereas high-income nations have some of the lowest blood pressure levels in the world[61]. Although historically, high as well as low blood pressure levels were observed in tropical countries[27], blood pressure levels seem to have increased in Africa and Asia over the past few decades, while declining sharply in high-income countries[61,62]. Of the cardiometabolic risk factors, only cholesterol reflects the classical epidemiological transition model, and is currently lower in tropical Africa, south Asia, and east and southeast Asia than in high-income Western countries (Fig. 5)[63].

Fetal and early-life environment and nutrition

Nutrition and environment during the fetal, childhood and adolescence periods are an important determinant of NCDs. In the second half of the twentieth century, researchers realised that children born or growing during periods of famine, war and economic hardship had

© 2018 Springer Nature Limited. All rights reserved.

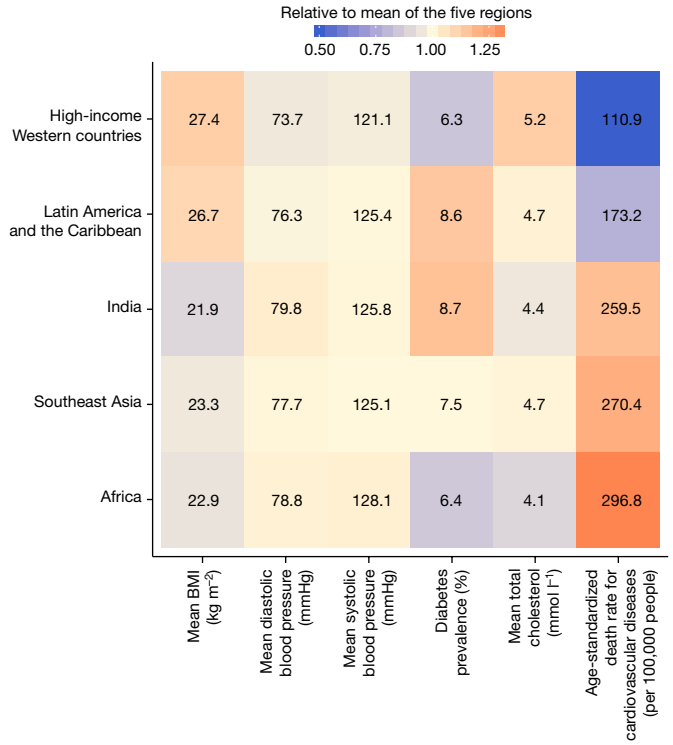


Fig. 5 | Age-standardized cardiovascular disease death rate, mean BMI, mean systolic and diastolic blood pressure, mean serum total cholesterol and diabetes prevalence in low- and middle-income tropical and the high-income Western region. For each outcome, the colour shows the ratio for the level in each region to the average of the five regions. The inset text gives the actual level of each outcome. For example, mean BMI in high-income Western countries is 27.4 kg m^{-2}, which is 1.12 times the average of all five regions. Data on deaths are from the WHO global health estimates[85] and on risk factors from the NCD Risk Factor Collaboration (http://www.ncdrisc.org)[44,45,61,63]. Death rates and risk factors were age-standardized to the WHO standard population[10].

a higher risk of some NCDs in adulthood[64–66]. The thrifty phenotype hypothesis[67], which summarises these findings, postulates that early-life undernutrition increases the risk of NCDs, especially if followed by rapid or excessive weight gain[68]. In parallel, nutritionists and social scientists noted that, although there is a strong genetic influence on individual height, changes in height of an entire population are driven by social, environmental and nutritional status during childhood and adolescence[69–71]. Being taller is associated with enhanced longevity, lower risk of cardiovascular and respiratory diseases, and higher risk of some cancers[65,72]. These studies collectively demonstrate that fetal, childhood and adolescent environment and nutrition influence NCDs, with at least part of the effects mediated by linear growth.

In the early twentieth century, the shortest heights in the world were those in tropical Asia and Andean Latin America[72]. Asians and Latin Americans grew taller over the course of the twentieth century, with some of the largest worldwide gains in height occurring in east Asia. In India and other south Asian nations, height in adulthood has plateaued at lower levels than those experienced elsewhere[72]. Adults in sub-Saharan Africa were taller than those in Asia and parts of Latin America in the early twentieth century, but fell behind because Africans born in the last three decades of the twentieth century are shorter than their predecessors[72,73]. Poor early-life nutrition and stunted growth may be partly responsible for the fact that some NCDs are high in low-income countries, and decline as economies develop, the opposite of what is predicted by a simple epidemiological transition model.

Infections, parasites and mycotoxins

A distinct feature of the NCDs in tropical countries, especially in the lower-income regions of tropical Africa and south Asia, is the importance of diseases with infectious, parasitic and fungal origins (Box 2). As discussed earlier, infection-related NCDs include rheumatic heart disease, liver cirrhosis, cervical, liver and stomach cancers, and Kaposi's sarcoma. Infections and parasitic diseases also contribute to other NCDs. For example, infectious respiratory diseases increase the risk of chronic obstructive pulmonary disease[74], schistosomiasis and other helminth infections increase the risk of genitourinary diseases[75], and inflammation

Box 2

Bidirectional links between infections and NCDs

Not only are infectious diseases an important cause of ill health in low- and middle-income countries, they also affect some NCDs, including many cancers[19,31,86]. That infections cause cancers was hypothesized centuries ago, and emerged as a subject of systematic research at the end of the nineteenth century, after Pasteur and Koch established the their role as a cause of disease in general[31,86]. Difficulties in establishing a clear link between specific infectious agents and cancer, together with the identification of ionizing radiation, soot, tar and other chemicals as causes of cancer, led to a lack of interest in, and even dismissal of, experimental results on infection and cancers in animals. After electron microscopy facilitated the discovery of the endogenous retroviruses, which could enter the host cell and modify its genome, interest in infectious, especially viral, causes of cancers revived. Researchers took advantage of new laboratory techniques and epidemiological studies in populations with high cancer rates to identify the causal roles of human T-cell leukaemia retrovirus (HTLV-1) in leukaemia, hepatitis B and C viruses in liver cancer, human papillomavirus in cervical cancer, Epstein-Barr virus in Burkitt's lymphoma, and HIV in Kaposi's sarcoma and non-Hodgkin's and Hodgkin's lymphoma[31]. Further research as well as chance observations (for example, high rates of acute liver disease in poultry fed with groundnuts) led to discovering non-viral communicable causes of cancers including bacteria, parasitic flukes and fungi[31]. The mechanisms for the carcinogenic effects of infections are varied and include oncogene activation, chronic inflammation, metaplasia of epithelium, release of carcinogens, general immunosuppression and loss of tumour suppressor ability[19].

The relationship between infections and NCDs is bidirectional. For example, tobacco smoking, alcohol use and diabetes not only affect NCDs, but also increase the risk of infectious diseases[87–91]. The established or hypothesized mechanisms for these effects include reduced immune function, such as through adverse impacts on macrophage and leukocyte function and lower production of interleukins; increased viral production and virulence in hyperglycaemic environments or after alcohol use; and lower adherence to anti-retroviral treatment and riskier sexual behaviour for alcohol and HIV[90,92]. Therefore, the rising prevalence of smoking, alcohol use and diabetes in low- and middle-income countries will set back efforts to combat classical tropical diseases such as tuberculosis and other respiratory infections[93–95].

The bidirectional relationship also provides an opportunity for integration of disease prevention, detection and treatment[96–98]. For example, screening tuberculosis patients can help to increase the currently low rate of diabetes diagnosis[98], and smoking and alcohol cessation and glycaemic control may improve the prognosis of tuberculosis patients. Direct evidence on the benefits of integration is currently lacking, especially from low- and middle-income countries with the highest infection burden, and will inevitably depend on extending health system access and utilization, health workforce training, availability of medicines for treatment, and patient follow-up to ensure treatment adherence.

© 2018 Springer Nature Limited. All rights reserved.

Box 3

Effective interventions for NCDs in low- and middle-income tropical countries

Reducing the burden of NCDs in low- and middle-income tropical countries requires a combination of prevention, early detection and treatment. We present here some interventions and policies to address NCDs that have a large burden in low- and middle-income countries (such as stroke, ischaemic heart disease and some cancers) as well as those predominant in such populations (such as cervical cancer, rheumatic heart disease and haemoglobinopathies). In selecting these strategies, we used studies and reports that had quantitatively or qualitatively analysed the expected impacts of health interventions, such as the Disease Control Priorities[49,99,100], the Lancet Series on NCDs[5], the Lancet Global Mental Health Series[101], and the Lancet Series on Maternal and Child Nutrition[102,103], as well as clinical guidelines such as those by the WHO and the National Institute for Health and Care Excellence. In relevant cases, such as for emergency response, we adapted the recommendations so that they can be implemented in settings with limited transportation and healthcare infrastructure.

Environment, nutrition and health behaviours

- Provide access to safe and clean housing, water, sanitation, electricity and refrigeration to reduce infectious and parasitic causes of NCDs.
- Implement public health programmes and social and agricultural policies that improve maternal, neonatal and child nutrition[102,103].
- Implement comprehensive tobacco and alcohol control, including taxation, warning labels, public smoking bans, and restriction of sales, marketing and advertising[49,104].
- Eliminate industrial transfats in food supply; reduce salt in processed foods; implement food policies that increase access to and consumption of fruits and vegetables; and tax and restrict the sales and marketing of sugar-sweetened beverages and other refined carbohydrates[100].

Healthcare

- Implement universal, financially and physically accessible, high-quality primary and specialist care systems, with workforce, infrastructure and medicines equipped to manage NCDs, and effective referral pathways from primary to specialist care[104,105].
- Implement and increase the coverage of school and community-based human papillomavirus and hepatitis B virus vaccination (including birth dose for hepatitis B virus to prevent mother-to-child transmission)[49].
- Increase access to and encourage cervical cancer screening for all women through opportunistic screening with visual inspection and outreach programmes for vulnerable groups, and treat precancerous lesions.
- Treatment of streptococcal pharyngitis with effective antibiotics to prevent acute rheumatic fever and spread of group A streptococcus.
- Measure blood pressure and use off-patent medicines to treat hypertension at primary care level, especially where laboratories for risk-based management of cardiovascular diseases are limited.
- Equip primary care for early detection of diabetes, treatable cancers, and kidney disease, and referral into effective treatment pathways.
- Implement newborn and paediatric NCD services for genetic and congenital disorders, type I diabetes, asthma and paediatric cancers in primary care, effective referral pathways to specialist care, treatment and management of such conditions and their complications, and patient and parent education and empowerment programmes.
- Use mobile phones and other digital technologies to create a decentralized emergency response system, train and equip first responders for stabilization and risk stratification of patients with possible myocardial infarction and stroke, and implement community care pathways for rapid transfer of such patients to a specialized healthcare facility.
- Implement pharmacological and psychosocial interventions for mental health conditions including depression and schizophrenia at the community and primary care level[101].

Research and development

- Develop an effective vaccine to prevent *Helicobacter pylori* infection.
- Develop robust and low-cost diagnostic tools and devices for detection of early-stage cancers in the community setting, as has been done for HIV and other infectious diseases.
- Characterize NCD and NCD risk factor phenotypes and traits—multi-morbidity with multiple communicable diseases and NCDs, and co-exposure to multiple environmental, infectious and physiological NCD risk factors—to identify optimal packages of interventions.
- Investigate, in epidemiological and health service studies, the aetiological relationship between infectious diseases and NCDs, and the feasibility and benefits of integrating their diagnoses and treatment.

as a result of chronic and repeated infections increases the risk of cardiovascular diseases[19,76].

Although infections remain an important cause of NCDs in poor countries and communities, the burden of most infection-related NCDs is comparable to or smaller than that of other NCDs. For example, colorectal cancer, which is caused by high consumption of red and processed meat and low intake of dietary fibre now has a comparable burden to liver and stomach cancers in low- and middle-income tropical regions (Fig. 4). Age-standardized death rates from breast cancer are higher than those of cervical cancer in every tropical region except Africa, where their mortality burden is about the same. Similarly, prostate cancer has emerged as an important cause of death among men across low- and middle-income tropical regions, with a higher age-standardized death rate than liver or stomach cancer in some regions. Finally, NCDs with non-infectious origins are not only an important cause of mortality in low- and middle-income countries, but they may increase the risk of infectious diseases (Box 2). This reciprocal relationship should motivate a more integrated response to communicable and non-communicable diseases.

Poverty

More than two-thirds of people in the poorest countries in tropical Africa and about 12% in India live on less than I$1.90 per day (I$, international dollar that is adjusted for purchasing power and inflation; https://data.worldbank.org/indicator/SI.POV.DDAY?). The poor experience higher rates of NCDs partly because many causes of NCDs, including poor nutrition at all stages of life; inadequate housing, water, sanitation and electricity; smoking and elevated blood pressure, tend to be associated with poverty[5,9,60,77–79]. People in poor countries and communities, and those with lower income, also have more limited access to high-quality healthcare, because health infrastructure, personnel

© 2018 Springer Nature Limited. All rights reserved.

and technologies are more restricted or because they cannot afford the out-of-pocket cost of services. Limited primary care leads to low rates of diagnosis and treatment, or late diagnosis. For example, fewer women are screened and treated for cervical cancer and fewer people with hypertension and diabetes are diagnosed and treated in lower-income than in high-income countries[5,80–82]. Late diagnosis, typically when complications arise, coupled with more limited specialist care, worsens the prognosis of NCDs in low- and middle-income countries compared to high-income nations[32,52,83]. Given its central role in pathogenesis of NCDs, the poverty of individuals and communities, and not inherent features of the tropics, is the most important determinant of the higher NCD rates in the lower-income tropical countries and regions, compared to high-income nations as well as to the better-off region of Latin America and the Caribbean.

Actions to reduce NCDs in low- and middle-income countries

The data presented in this Review demonstrate that NCDs are an important cause of ill health and early death in low- and middle-income tropical countries, and after accounting for differences in age, most NCDs have a higher mortality burden than in high-income Western countries. Therefore, reducing the burden of NCDs is essential for improving health in low- and middle-income tropical countries and reducing global health inequalities.

The causes of NCDs in low- and middle-income tropical countries are diverse and include poor nutrition and living environment, infections, insufficient regulation of tobacco and alcohol, and under-resourced and inaccessible healthcare. Although much of the epidemiological research has focused on the effect of these factors on individual NCDs, they often lead to multiple communicable and NCDs in the same people and communities, leading to multi-morbidity[84]. Multiple co-occurring diseases and risk factors can have amplifying effects and lead to overall poor health and loss of productivity. At the same time, the shared causes and aetiologies of these conditions, and their overlap, provides an opportunity for systematic integration of public health and healthcare interventions for their prevention and treatment. Doing so requires better data on the patterns of multi-morbidity among communicable and non-communicable diseases and their risk factors in low- and middle-income countries because they may be different from those in high-income countries (Box 3).

Underpinning these proximal determinants is a background of poverty, food insecurity (in terms of both quantity and quality), and insufficiency of infrastructure and services related to housing, water, sanitation, transportation and healthcare that afflict poor countries, communities and families. These challenges are the focus of the SDGs. The SDGs have broadened the development agenda of their predecessor, the Millennium Development Goals, whose health-related goals and targets focused on maternal and child health and infectious diseases. As stated earlier, SDGs have a target related to reducing NCD mortality, although this target excludes some important NCD causes of death in low- and middle-income tropical countries such as kidney, liver and other genitourinary and digestive diseases. Alongside this NCD mortality target, SDGs have targets for reducing tobacco and alcohol use (important causes of NCDs throughout the world), and for increasing universal health coverage (essential for enhancing timely diagnosis and treatment of NCDs and their risk factors in an equitable manner). The momentum generated by the SDGs provides an opportunity to turn political attention to NCDs. To succeed in reducing the large burden of NCDs in low- and middle-income countries and global health inequalities requires (1) identifying interventions and programmes that can be scaled up in resource-limited settings to effectively lower the burden of NCDs (Box 3) and (2) a substantial increase in domestic as well as international financing that supports low- and middle-income countries in sustained actions against NCDs.

Received: 10 January 2018; Accepted: 24 May 2018;
Published online 25 July 2018.

1. Wagner, K. H. & Brath, H. A global view on the development of non communicable diseases. *Prev. Med.* **54**, S38–S41 (2012).
2. McKeown, T. *The Origins of Human Disease* (Blackwell, Oxford, 1988).
3. Omran, A. R. The epidemiologic transition. A theory of the epidemiology of population change. *Milbank Mem. Fund Q.* **49**, 509–538 (1971).
4. Nugent, R. A chronology of global assistance funding for NCD. *Glob. Heart* **11**, 371–374 (2016).
5. Di Cesare, M. et al. Inequalities in non-communicable diseases and effective responses. *Lancet* **381**, 585–597 (2013).
A synthesis of data on global and within-country inequalities in NCDs and their causes, showing that the burden of ill health from most NCDs is higher in poor countries, poor communities and poor people.
6. Lopez, A. D., Mathers, C. D., Ezzati, M., Jamison, D. T. & Murray, C. J. Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data. *Lancet* **367**, 1747–1757 (2006).
7. World Health Organization (WHO). *Global Status Report on Noncommunicable Diseases 2010*. http://www.who.int/nmh/publications/ncd_report2010/en/ (WHO, 2011).
8. Bray, F., Jemal, A., Grey, N., Ferlay, J. & Forman, D. Global cancer transitions according to the Human Development Index (2008-2030): a population-based study. *Lancet Oncol.* **13**, 790–801 (2012).
9. Kwan, G. F. et al. Endemic cardiovascular diseases of the poorest billion. *Circulation* **133**, 2561–2575 (2016).
10. Ahmad, O. et al. *Age Standardization of Rates: a New WHO Standard*. http://www.who.int/healthinfo/paper31.pdf (World Health Organization, 2001).
11. Moriyama, I. M. & Gover, M. Heart diseases and allied causes of death in relation to age changes in the population. *Public Health Rep.* **63**, 537–545 (1948).
An elegant and pioneering analysis of trends over time in medical causes of death that shows the diversity of trends in deaths from different NCDs.
12. Preston, S. H. & Nelson, V. E. Structure and change in causes of death: an international summary. *Popul. Stud.* **28**, 19–51 (1974).
13. Omran, A. R. A century of epidemiologic transition in the United States. *Prev. Med.* **6**, 30–51 (1977).
14. Mayosi, B. M. in *Braunwald's Heart Disease: A Textbook of Cardiovascular Medicine.* 10th edn (eds Mann, D. L., Zipes, D. P., Libby, P. & Bonow, R. O.) (Elsevier Saunders, Philadelphia 2015).
15. Gordis, L. The virtual disappearance of rheumatic fever in the United States: lessons in the rise and fall of disease. T. Duckett Jones memorial lecture. *Circulation* **72**, 1155–1162 (1985).
16. Ezzati, M. et al. Contributions of risk factors and medical care to cardiovascular mortality trends. *Nat. Rev. Cardiol.* **12**, 508–530 (2015).
17. Karthikeyan, G. & Mayosi, B. M. Is primary prevention of rheumatic fever the missing link in the control of rheumatic heart disease in Africa? *Circulation* **120**, 709–713 (2009).
18. Shaper, A. G. Cardiovascular disease in the tropics. II. Endomyocardial fibrosis. *BMJ* **3**, 743–746 (1972).
19. Ogoina, D. & Onyemelukwe, G. C. The role of infections in the emergence of non-communicable diseases (NCDs): compelling needs for novel strategies in the developing world. *J. Infect. Public Health* **2**, 14–29 (2009).
20. Global Burden of Metabolic Risk Factors for Chronic Diseases Collaboration. Cardiovascular disease, chronic kidney disease, and diabetes mortality burden of cardiometabolic risk factors from 1980 to 2010: a comparative risk assessment. *Lancet Diabetes Endocrinol.* **2**, 634–647 (2014).
21. Gladwin, M. T. Cardiovascular complications and risk of death in sickle-cell disease. *Lancet* **387**, 2565–2574 (2016).
22. Woolsey, T. D. & Moriyama, I. M. Statistical studies of heart diseases; important factors in heart disease mortality trends. *Public Health Rep.* **63**, 1247–1273 (1948).
23. Shaper, A. G. Cardiovascular studies in the Samburu tribe of Northern Kenya. *Am. Heart J.* **63**, 437–442 (1962).
24. Shaper, A. G. Cardiovascular disease in the tropics. IV. Coronary heart disease. *BMJ* **4**, 32–35 (1972).
One of a series of four articles that presented carefully collected high-quality data on cardiovascular diseases in low-income tropical countries, especially in Africa, and demonstrated the complexity of their aetiology.
25. Kennelly, B. M., Truswell, A. S. & Schrire, V. A clinical and electrocardiographic study of Kung Bushmen. *S. Afr. Med. J.* **46**, 1093–1097 (1972).
26. Kaplan, H. et al. Coronary atherosclerosis in indigenous South American Tsimane: a cross-sectional cohort study. *Lancet* **389**, 1730–1739 (2017).
27. Shaper, A. G. Problems and prospects in tropical cardiology. *Scott. Med. J.* **12**, 393–400 (1967).
28. Wild, C. P. The role of cancer research in noncommunicable disease control. *J. Natl. Cancer Inst.* **104**, 1051–1058 (2012).
29. Danaei, G., Vander Hoorn, S., Lopez, A. D., Murray, C. J. & Ezzati, M. Causes of cancer in the world: comparative risk assessment of nine behavioural and environmental risk factors. *Lancet* **366**, 1784–1793 (2005).
30. Ezzati, M. & Riboli, E. Can noncommunicable diseases be prevented? Lessons from studies of populations and individuals. *Science* **337**, 1482–1487 (2012).
31. Blackadar, C. B. Historical review of the causes of cancer. *World J. Clin. Oncol.* **7**, 54–86 (2016).
32. Allemani, C. et al. Global surveillance of cancer survival 1995-2009: analysis of individual data for 25,676,887 patients from 279 population-based registries in 67 countries (CONCORD-2). *Lancet* **385**, 977–1010 (2015).
33. Plummer, M. et al. Global burden of cancers attributable to infections in 2012: a synthetic analysis. *Lancet Glob. Health* **4**, e609–e616 (2016).
34. Marshall, B. J. & Warren, J. R. Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. *Lancet* **1**, 1311–1315 (1984).

© 2018 Springer Nature Limited. All rights reserved.

The seminal study that established the role of Helicobacter pylori as a cause of lesions that precede stomach cancer, and earned Marshall and Warren the Nobel Prize in Physiology and Medicine.
35. Gersten, O. & Wilmoth, J. R. The cancer transition in Japan since 1951. *Demogr. Res.* **7**, 271–306 (2002).
36. Wingo, P. A. et al. Long-term trends in cancer mortality in the United States, 1930-1998. *Cancer* **97**, 3133–3275 (2003).
37. Howson, C. P., Hiyama, T. & Wynder, E. L. The decline in gastric cancer: epidemiology of an unplanned triumph. *Epidemiol. Rev.* **8**, 1–27 (1986).
38. Vaccarella, S., Lortet-Tieulent, J., Plummer, M., Franceschi, S. & Bray, F. Worldwide trends in cervical cancer incidence: impact of screening against changes in disease risk factors. *Eur. J. Cancer* **49**, 3262–3273 (2013).
39. Ferro, A. et al. Worldwide trends in gastric cancer mortality (1980-2011), with predictions to 2015, and incidence by subtype. *Eur. J. Cancer* **50**, 1330–1344 (2014).
40. Bruni, L. et al. Global estimates of human papillomavirus vaccination coverage by region and income level: a pooled analysis. *Lancet Glob. Health* **4**, e453–e463 (2016).
41. Xia, J., Jiang, S. C. & Peng, H. J. Association between liver fluke infection and hepatobiliary pathological changes: a systematic review and meta-analysis. *PLoS One* **10**, e0132673 (2015).
42. World Health Organization (WHO). *Global Hepatitis Report 2017*. (http://www.who.int/hepatitis/publications/global-hepatitis-report2017/en/ (WHO, 2017).
43. Pearson-Stuttard, J. et al. Worldwide burden of cancer attributable to diabetes and high body-mass index: a comparative risk assessment. *Lancet Diabetes Endocrinol.* **6**, e6–e15 (2018).
44. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants. *Lancet* **387**, 1513–1530 (2016).
45. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128·9 million children, adolescents, and adults. *Lancet* **390**, 2627–2642 (2017).
An analysis of trends in underweight, overweight and obesity for all of the world's countries using the largest-ever collection anthropometric measurement from nearly 130 million people over four decades.
46. Ezzati, M. & Riboli, E. Behavioral and dietary risk factors for noncommunicable diseases. *N. Engl. J. Med.* **369**, 954–964 (2013).
47. Boseley, S. *Threats, bullying, lawsuits: tobacco industry's dirty war for the African market. The Guardian* https://www.theguardian.com/world/2017/jul/12/big-tobacco-dirty-war-africa-market (12 July 2017).
48. World Health Organization (WHO). *Global Status Report on Alcohol and Health 2014* (WHO, 2014).
49. Gelband, H. et al. Costs, affordability, and feasibility of an essential package of cancer control interventions in low-income and middle-income countries: key messages from *Disease Control Priorities*, 3rd edition. *Lancet* **387**, 2133–2144 (2016).
50. Weatherall, D. J. & Clegg, J. B. Inherited haemoglobin disorders: an increasing global health problem. *Bull. World Health Organ.* **79**, 704–712 (2001).
51. Piel, F. B. et al. Global distribution of the sickle cell gene and geographical confirmation of the malaria hypothesis. *Nat. Commun.* **1**, 104 (2010).
A careful analysis that confirmed the role of historical malaria geography as the determinant of the geographical distribution of sickle cell disease.
52. Grosse, S. D. et al. Sickle cell disease in Africa: a neglected cause of early childhood mortality. *Am. J. Prev. Med.* **41**, S398–S405 (2011).
53. McGinley, K. F., Tay, K. J. & Moul, J. W. Prostate cancer in men of African origin. *Nat. Rev. Urol.* **13**, 99–107 (2016).
54. Haiman, C. A. et al. Genome-wide association study of prostate cancer in men of African ancestry identifies a susceptibility locus at 17q21. *Nat. Genet.* **43**, 570–573 (2011).
55. Popkin, B. M. Global nutrition dynamics: the world is shifting rapidly toward a diet linked with noncommunicable diseases. *Am. J. Clin. Nutr.* **84**, 289–298 (2006).
56. Reddy, K. S. & Yusuf, S. Emerging epidemic of cardiovascular disease in developing countries. *Circulation* **97**, 596–601 (1998).
57. World Health Organization (WHO). *Global Report on Urban Health: Equitable Healthier Cities for Sustainable Development*. Report No. 9241565276 (World Health Organization, Geneva, 2016).
58. Yusuf, S., Reddy, S., Ôunpuu, S. & Anand, S. Global burden of cardiovascular diseases: part I: general considerations, the epidemiologic transition, risk factors, and impact of urbanization. *Circulation* **104**, 2746–2753 (2001).
59. Micha, R. et al. Global, regional and national consumption of major food groups in 1990 and 2010: a systematic analysis including 266 country-specific nutrition surveys worldwide. *BMJ Open* **5**, e008705 (2015).
60. Danaei, G. et al. The global cardiovascular risk transition: associations of four metabolic risk factors with national income, urbanization, and Western diet in 1980 and 2008. *Circulation* **127**, 1493–1502 (2013).
61. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in blood pressure from 1975 to 2015: a pooled analysis of 1479 population-based measurement studies with 19.1 million participants. *Lancet* **389**, 37–55 (2017).
An analysis of four decades of trends in systolic and diastolic blood pressure for all of the world's countries that showed that low-income countries have some of the highest blood pressure levels in the world, and vice versa.
62. NCD Risk Factor Collaboration (NCD-RisC). Contributions of mean and shape of blood pressure distribution to worldwide trends and variations in raised blood pressure: a pooled analysis of 1018 population-based measurement studies with 88.6 million participants. *Int. J. Epidemiol.* **47**, 872–883i (2018).
63. Farzadfar, F. et al. National, regional, and global trends in serum total cholesterol since 1980: systematic analysis of health examination surveys and epidemiological studies with 321 country-years and 3.0 million participants. *Lancet* **377**, 578–586 (2011).
64. Barker, D. J., Winter, P. D., Osmond, C., Margetts, B. & Simmonds, S. J. Weight in infancy and death from ischaemic heart disease. *Lancet* **2**, 577–580 (1989).
65. Victora, C. G. et al. Maternal and child undernutrition: consequences for adult health and human capital. *Lancet* **371**, 340–357 (2008).
66. Ravelli, G. P., Stein, Z. A. & Susser, M. W. Obesity in young men after famine exposure in utero and early infancy. *N. Engl. J. Med.* **295**, 349–353 (1976).
One of the first studies to show the role of fetal and early life undernutrition as a cause of NCDs and NCD risk factors in adult life.
67. Hales, C. N. & Barker, D. J. Type 2 (non-insulin-dependent) diabetes mellitus: the thrifty phenotype hypothesis. *Diabetologia* **35**, 595–601 (1992).
68. Bhargava, S. K. et al. Relation of serial changes in childhood body-mass index to impaired glucose tolerance in young adulthood. *N. Engl. J. Med.* **350**, 865–875 (2004).
69. Tanner, J. M. Growth as a mirror of the condition of society: secular trends and class distinctions. *Acta Paediatr. Jpn.* **29**, 96–103 (1987).
70. Tanner, J. M. *A History of the Study of Human Growth* (Cambridge University Press, Cambridge, 1981).
71. Fogel, R. W. *Changes in the Physiology of Aging during the Twentieth Century.* NBER Working Paper 11233 (National Bureau of Economic Research, 2005).
72. NCD Risk Factor Collaboration (NCD-RisC). A century of trends in adult human height. *eLife* **5**, e13410 (2016).
The first analysis of trends in height in adulthood for all of the countries in the world, that showed the massive worldwide variation in height gain over a period of a century.
73. Stevens, G. A. et al. Trends in mild, moderate, and severe stunting and underweight, and progress towards MDG 1 in 141 developing countries: a systematic analysis of population representative data. *Lancet* **380**, 824–834 (2012).
74. Byrne, A. L., Marais, B. J., Mitnick, C. D., Lecca, L. & Marks, G. B. Tuberculosis and chronic respiratory disease: a systematic review. *Int. J. Inf. Dis.* **32**, 138–146 (2015).
75. Colley, D. G., Bustinduy, A. L., Secor, W. E. & King, C. H. Human schistosomiasis. *Lancet* **383**, 2253–2264 (2014).
76. Roivainen, M. et al. Infections, inflammation, and the risk of coronary heart disease. *Circulation* **101**, 252–257 (2000).
77. Black, R. E. et al. Maternal and child undernutrition and overweight in low-income and middle-income countries. *Lancet* **382**, 427–451 (2013).
78. Gakidou, E. et al. Improving child survival through environmental and nutritional interventions: the importance of targeting interventions toward the poor. *J. Am. Med. Assoc.* **298**, 1876–1887 (2007).
79. Shaper, A. G. Cardiovascular disease in the tropics. I. Rheumatic heart. *BMJ* **3**, 683–686 (1972).
80. Gakidou, E., Nordhagen, S. & Obermeyer, Z. Coverage of cervical cancer screening in 57 countries: low average levels and large inequalities. *PLoS Med.* **5**, e132 (2008).
81. Chow, C. K. et al. Prevalence, awareness, treatment, and control of hypertension in rural and urban communities in high-, middle-, and low-income countries. *J. Am. Med. Assoc.* **310**, 959–968 (2013).
82. Mills, K. T. et al. Global disparities of hypertension prevalence and control: a systematic analysis of population-based studies from 90 countries. *Circulation* **134**, 441–450 (2016).
83. Feigin, V. L., Lawes, C. M., Bennett, D. A., Barker-Collo, S. L. & Parag, V. Worldwide stroke incidence and early case fatality reported in 56 population-based studies: a systematic review. *Lancet Neurol.* **8**, 355–369 (2009).
84. Academy of Medical Sciences (AMA). *Multimorbidity: a Priority for Global Health Research*. https://acmedsci.ac.uk/policy/policy-projects/multimorbidity (AMA, 2018).
85. World Health Organization (WHO). *Global Health Estimates: Deaths by Cause, Age, Sex and Country, 2000–2016* http://www.who.int/healthinfo/global_burden_disease/estimates/en/ (WHO, 2018).
86. Epstein, M. A. Historical background. *Phil. Trans. R. Soc. Lond. B* **356**, 413–420 (2001).
87. Bagaitkar, J., Demuth, D. R. & Scott, D. A. Tobacco use increases susceptibility to bacterial infection. *Tob. Induc. Dis.* **4**, 12 (2008).
88. Lin, H. H., Ezzati, M. & Murray, M. Tobacco smoke, indoor air pollution and tuberculosis: a systematic review and meta-analysis. *PLoS Med.* **4**, e20 (2007).
89. Parry, C., Rehm, J., Poznyak, V. & Room, R. Alcohol and infectious diseases: an overlooked causal linkage? *Addiction* **104**, 331–332 (2009).
90. Rehm, J., Probst, C., Shield, K. D. & Shuper, P. A. Does alcohol use have a causal effect on HIV incidence and disease progression? A review of the literature and a modeling strategy for quantifying the effect. *Popul. Health Metr.* **15**, 4 (2017).
91. Casqueiro, J., Casqueiro, J. & Alves, C. Infections in patients with diabetes mellitus: A review of pathogenesis. *Indian J. Endocrinol. Metab.* **16**, S27–S36 (2012).
92. Pandrea, I., Happel, K. I., Amedee, A. M., Bagby, G. J. & Nelson, S. Alcohol's role in HIV transmission and disease progression. *Alcohol Res. Health* **33**, 203–218 (2010).
93. Imtiaz, S. et al. Alcohol consumption as a risk factor for tuberculosis: meta-analyses and burden of disease. *Eur. Respir. J.* **50**, 1700216 (2017).
94. Pan, S. C. et al. Effect of diabetes on tuberculosis control in 13 countries with high tuberculosis: a modelling study. *Lancet Diabetes Endocrinol.* **3**, 323–330 (2015).

© 2018 Springer Nature Limited. All rights reserved.

95. Harries, A. D. et al. The looming epidemic of diabetes-associated tuberculosis: learning lessons from HIV-associated tuberculosis. *Int. J. Tuberc. Lung Dis.* **15**, 1436–1444 (2011).
96. Dye, C. After 2015: infectious diseases in a new era of health and development. *Phil. Trans. R. Soc. Lond. B* **369**, 20130426 (2014).
97. Baris, E. & Ezzati, M. Should interventions to reduce respirable pollutants be linked to tuberculosis control programmes? *Br. Med. J.* **329**, 1090–1093 (2004).
98. Harries, A. D. et al. Diabetes mellitus and tuberculosis: programmatic management issues. *Int. J. Tuberc. Lung Dis.* **19**, 879–886 (2015).
99. Jamison, D. T. et al. Global health 2035: a world converging within a generation. *Lancet* **382**, 1898–1955 (2013).
100. Prabhakaran, D. et al. Cardiovascular, respiratory, and related disorders: key messages from *Disease Control Priorities*, 3rd edition. *Lancet* **391**, 1224–1236 (2018).
101. Patel, V. et al. Treatment and prevention of mental disorders in low-income and middle-income countries. *Lancet* **370**, 991–1005 (2007).
102. Ruel, M. T. & Alderman, H. Nutrition-sensitive interventions and programmes: how can they help to accelerate progress in improving maternal and child nutrition? *Lancet* **382**, 536–551 (2013).
103. Bhutta, Z. A. et al. Evidence-based interventions for improvement of maternal and child nutrition: what can be done and at what cost? *Lancet* **382**, 452–477 (2013).
104. Jamison, D. T. et al. Universal health coverage and intersectoral action for health: key messages from *Disease Control Priorities*, 3rd edition. *Lancet* **391**, 1108–1120 (2018).
The overview of the work in the Disease Control Priorities study that identified actions and interventions with potential for large impacts on different diseases, including NCDs.
105. Manjomo, R. C. et al. Managing and monitoring chronic non-communicable diseases in a primary health care clinic, Lilongwe, Malawi. *Public Health Action* **6**, 60–65 (2016).

Acknowledgements We thank R. Beaglehole, S. Blundell, R. Nugent, F. Piel, M. Parkins and M. Thun for comments and recommendation of background literature.

Reviewer information *Nature* thanks F. Assah, P. Byass, B. Singer and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions M.E. designed the Review concept and wrote the Review with input from J.P.-S., J.E.B. and C.D.M. J.P.-S. contributed to clinical and epidemiological reviews. J.E.B. and C.D.M. analysed data, and J.E.B. prepared figures.

Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0306-9.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.E.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

REVIEW

https://doi.org/10.1038/s41586-018-0301-1

The future of hyperdiverse tropical ecosystems

Jos Barlow[1]*, Filipe França[1,2], Toby A. Gardner[3], Christina C. Hicks[1], Gareth D. Lennox[1], Erika Berenguer[1,4], Leandro Castello[5], Evan P. Economo[6], Joice Ferreira[2], Benoit Guénard[7], Cecília Gontijo Leal[8], Victoria Isaac[9], Alexander C. Lees[10], Catherine L. Parr[11,12,13], Shaun K. Wilson[14,15], Paul J. Young[1] & Nicholas A. J. Graham[1]

The tropics contain the overwhelming majority of Earth's biodiversity: their terrestrial, freshwater and marine ecosystems hold more than three-quarters of all species, including almost all shallow-water corals and over 90% of terrestrial birds. However, tropical ecosystems are also subject to pervasive and interacting stressors, such as deforestation, overfishing and climate change, and they are set within a socio-economic context that includes growing pressure from an increasingly globalized world, larger and more affluent tropical populations, and weak governance and response capacities. Concerted local, national and international actions are urgently required to prevent a collapse of tropical biodiversity.

The tropics hold a disproportionate amount of global biological diversity and are key to meeting the international community's aims of socially just sustainable development and effective biodiversity conservation[1]. Yet, tropical ecosystems are undergoing rapid environmental, socio-economic and demographic change[2], often driven by forces originating in extra-tropical developed countries. The scale of these changes is unprecedented, and decisions implemented in the coming decades will define the future sustainability of the tropics.

Guiding these decisions depends on understanding the diversity and vulnerability of the four major tropical ecosystems: the forests and mesic savannahs that cover most of the terrestrial tropics, the extensive freshwater systems that receive half of the world's rainfall and the shallow-water coral reefs distributed along 150,000 km of coastline (Fig. 1). Here we quantify and review the global importance of tropical biodiversity, evaluate the vulnerability of tropical ecosystems to proximate stressors and assess whether global and regional socio-economic changes will exacerbate or ameliorate biodiversity loss. We then examine the effectiveness of conservation approaches and highlight the scientific advances required to support a sustainable tropical future.

The global importance of tropical ecosystems

Over evolutionary time, the tropics have acted as both a source of and a refuge for extra-tropical terrestrial and marine species[3,4]; but just how diverse and irreplaceable are the tropics today? The increase in species richness from polar to tropical regions, known as the latitudinal diversity gradient, is found across a wide range of taxa and biomes. As a result of this gradient, tropical latitudes—which cover just 40% of the Earth's surface—hold a startling proportion of the planet's species: our assessment reveals that almost all shallow-water zooxanthellae corals, 91% of terrestrial birds, and over 75% of amphibians, terrestrial mammals, freshwater fish, ants, flowering plants and marine fish have ranges that intersect tropical latitudes (Fig. 2a). For birds, the importance of the tropics extends far beyond 23.5 degrees of latitude, given that almost half of all Nearctic species migrate to the Neotropics[5] and over 2 billion individual passerines and near-passerines cross the Sahara each autumn[6]. Moreover, a disproportionate number of species are endemic to the tropics. For example, there are more than six times as many endemic terrestrial bird species in the tropics as in temperate regions (Fig. 2a). Tropical zones are less important for marine mammals and birds, which peak in diversity at mid-latitudes[7,8]. Nonetheless, more than 55% of these species use the tropics (Fig. 2a).

Overall, 78% of species across the ten taxa that we assessed occurred within tropical latitudes, but incomplete taxonomic inventories mean that this is almost certainly an underestimate[9]. Between 15,000 and 19,000 new species are described annually[10], and the majority of recently described terrestrial vertebrates[11] or predicted discoveries of invertebrates[12] are from the tropics. Even terrestrial mammals are still being discovered at a rate of about 25 species a year, with the highest numbers in the Neo- and Afrotropics[13]. Shortfalls in species descriptions for other taxa are often far greater. For example, only 70,000 of an estimated 830,000 multicellular plants and animals have been named on coral reefs[14], and although approximately 500 spider species are described each year, this is a tiny fraction of the estimated 150,000 undescribed tropical species[15].

Tropical taxonomic shortfalls are further compounded by a suite of systematic sampling biases. These include undersampling when compared with temperate regions[16], the spatial aggregation of sampling effort around coastal areas[17], roads, rivers, urban settlements and high-profile research stations[18], biases in favour of dry-season sampling when many invertebrate taxa are least abundant[19], and the paucity of samples from ecosystems that are harder to access, such as mesophotic and rariphotic reefs[20].

The biological diversity of the tropics is mirrored by many forms of societal diversity[21]. For example, tropical countries contain 40% of the world's population yet 85% of extant languages are spoken within them[22]. The tropics also provide incalculable benefits to humanity. They housed most of the key centres of plant domestication[23] and have been a vital laboratory for the development of science itself—the disciplines of ecology, biogeography and evolutionary biology are founded on evidence gleaned from tropical ecosystems. Tropical ecosystems also

[1]Lancaster Environment Centre, Lancaster University, Lancaster, UK. [2]Embrapa Amazônia Oriental, Belém, Brazil. [3]Stockholm Environment Institute, Stockholm, Sweden. [4]Environmental Change Institute, University of Oxford, Oxford, UK. [5]Department of Fish and Wildlife Conservation, Virginia Tech, Blacksburg, VA, USA. [6]Biodiversity and Biocomplexity Unit, Okinawa Institute of Science and Technology Graduate University, Onna, Japan. [7]School of Biological Sciences, The University of Hong Kong, Hong Kong, China. [8]Museu Paraense Emílio Goeldi, Belém, Brazil. [9]Universidade Federal do Pará, Belém, Brazil. [10]School of Science and the Environment, Manchester Metropolitan University, Manchester, UK. [11]School of Environmental Sciences, University of Liverpool, Liverpool, UK. [12]University of Pretoria, Pretoria, South Africa. [13]University of Witwatersrand, Johannesburg, South Africa. [14]Marine Science Program, Department of Biodiversity, Conservation and Attractions, Kensington, Western, Australia. [15]Oceans Institute, University of Western Australia, Crawley, Western Australia, Australia. *e-mail: josbarlow@gmail.com

© 2018 Springer Nature Limited. All rights reserved.

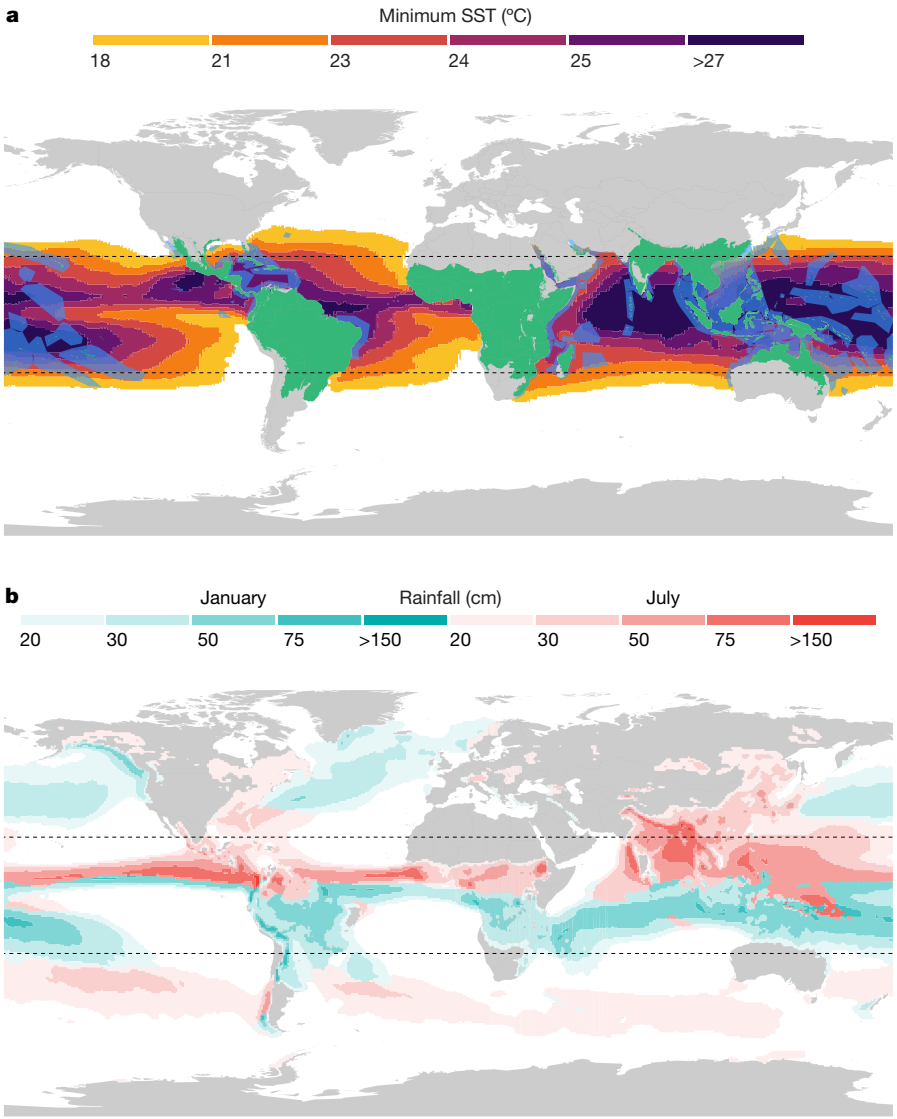


Fig. 1 | The tropical biosphere. a, Tropical terrestrial and marine biomes. The tropical terrestrial biome (green) was defined as all tropical mesic ecoregions[163]. These ecoregions span 82% of the 50 million km^2 of land between 23.5° N and 23.5° S, but extend into the subtropics in some areas. The tropical marine biome was defined by the 1988–2018 mean minimum monthly 18 °C sea-surface isotherm. This isotherm bounds the latitudinal extent of shallow-water coral-forming ecoregions (blue)[164]. **b**, The intertropical convergence zone (ITCZ). The ITCZ was defined as regions that received a mid-summer (January (turquoise colour gradient) or July (red colour gradient)) mean monthly total rainfall of >20 cm, for the period 1979–2017. Where both January and July had rainfall of >20 cm, we show the measurement from the month with the largest total. The ITCZ is a strong predictor of the distribution of the tropical terrestrial ecoregions shown in **a**. Data sources are presented in Supplementary Table 1.

make vital contributions to globally important ecosystem services. Although they cover just 0.1% of the ocean surface, coral reefs provide fish resources for the 275 million people that live within 30 km of them[24] and coastal protection for up to 197 million people[25]. Humid tropical forests cover less than 12% of the world's ice-free land surface but produce 33% of global net primary productivity and store 25% of the carbon in the terrestrial biosphere[26], and tropical savannahs provide a further 30% of global net primary productivity and 15% of carbon storage[27]. Tropical ecosystems also help drive vital atmospheric teleconnections. For instance, 70% of the rainfall in the 3.2-million-km^2 Rio de la Plata catchment is estimated to come from evaporation in Amazonia[28].

Vulnerability of tropical biodiversity

For each of the five vertebrate groups that have been comprehensively assessed by the International Union for the Conservation of Nature (IUCN) and for which spatial occurrence data are available[29], species classified as Vulnerable, Endangered or Critically Endangered are more dependent on the tropics than are those classified as Least Concern (Fig. 2b). In addition, 85% of species extinctions from these vertebrate groups have been of species that use the tropics[29]. Consequently, although extinctions of other groups are less well understood, we can assume that most of an estimated 130,000 modern invertebrate extinctions[30] will also have been of tropical species. Thus, not only are the tropics vastly more diverse than temperate regions, but this diversity is also at far greater risk from human actions[31]. Moreover, given that the tropics have the highest proportion of species classified by the IUCN as Data Deficient and the lowest level of biodiversity-threat assessment[16], information shortfalls mean we are probably underestimating the vulnerability of the tropical biome. We assessed this vulnerability in more depth by examining the effect of local and global stressors, the interactions between them and the resulting changes to tropical ecosystems.

Local stressors

The tropics are subject to some of the highest global rates of land-use change and degradation. Since 1990, while the spatial coverage of temperate forests has increased, tropical deforestation rates have

© 2018 Springer Nature Limited. All rights reserved.

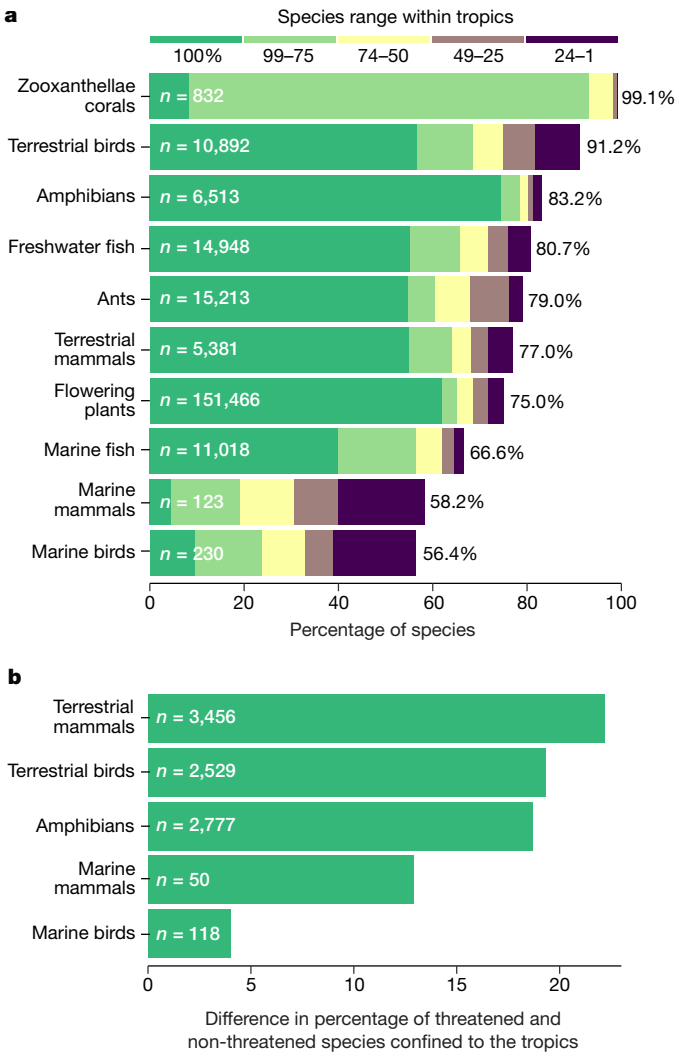


Fig. 2 | Tropical hyperdiversity. a, The proportion of species found within tropical latitudes for ten taxonomic groups. Bars are colour-coded to show the percentage of species ranges within the tropics. *n* gives the total number of species analysed in each group. Only birds, amphibians and mammals have been comprehensively sampled. Numbers at the end of the bars give the precise percentage of species whose ranges overlap tropical latitudes, as shown in the bars. **b**, The difference in the proportion of threatened (IUCN Critically Endangered, Endangered and Vulnerable categories) and non-threatened (IUCN Least Concern category) species found exclusively within tropical latitudes for the five comprehensively sampled groups. For example, 66% of threatened and 44% of non-threatened terrestrial mammals are confined to the tropics. The bars show the difference between these figures, which is 22% in the case of terrestrial mammals. Data are from Birdlife International[165] (for birds), the IUCN[29] (for amphibians and mammals), the Ocean Biogeographic Information System (for marine fish), C. Veron (personal communication) (for shallow-water zooxanthellate corals), the Global Biodiversity Information Facility (for flowering plants) and a previous publication[166] (for freshwater fish). Data sources are presented in Supplementary Table 1.

exceeded five million hectares per year[32]. Additional effects stem from the expansion of large infrastructure projects, such as dams, and the growing demand for agricultural commodities, biofuels, timber, wood for fuel, and other natural resources[33]. All of these result in severe biotic responses. Even with mitigation, dams present a near-impassable barrier for river fish[34], and deforestation replaces a species-rich pool of forest specialists with a smaller pool of common open-area species[35]. The influence of land-use change also extends far into remaining natural areas through isolation and edge effects[36], additional anthropogenic disturbances[37] and altered climatic conditions[38]. Edge effects suppress the abundance of threatened vertebrates up to 200–400 m into tropical forests[36]; this has left almost no core forest refugia in the Brazilian Atlantic forest, of which over 80% is within 500 m of an edge[39]. Even low levels of landscape modification have marked effects on range-restricted species[37], and time lags mean that some of the most deleterious outcomes are observed only decades after landscape modification[40].

Pollution presents a diverse set of threats to tropical ecosystems. Inputs of sediment and nutrients from land-use change are well-established drivers of biodiversity loss across freshwater[41] and coastal systems, including coral reefs[42]. Pesticide use is increasing across the tropics, reflecting rapid intensification of farming practices[43] and high pressures from pests on tropical crops[44]. Tropical Asian rivers are a major source of the 1.2–2.4 million tonnes of plastic that enters the world's oceans each year[45], with micro-plastics entering coral diets[46] and larger debris increasing rates of coral disease[47]. These examples of chronic pollution are exacerbated by extreme events such as the collapse of the Fundão Dam, which released about 50 million cubic metres of waste into a 600-km stretch of river in south-east Brazil and caused a 7,000-km^2 toxic plume in the Atlantic Ocean[48].

Overexploitation is also pervasive across the tropics. Fishing has reduced fish biomass by over 75% across a third of coral reefs[49] and is shrinking the mean body size of exploited freshwater taxa[50]. Hunting has contributed to the loss of charismatic megafauna, extirpating African elephants, rhinoceroses and large predators from most of their original ranges[51,52]. Tropical forests are affected by extensive over-harvesting of wildlife[31], with estimates of the annual harvests of highly trafficked animals such as pangolins reaching into the millions of individuals[53]. Moreover, the growth in non-food uses of wildlife means that even small-bodied songbirds are at risk of global extinction[54]. Overexploitation also extends beyond fauna and is driving economically valuable tropical tree species to extinction[55].

Invasive species have been the second most important driver of vertebrate extinctions since AD 1500[56]. Within terrestrial ecosystems, invasive species have exerted the strongest influence on islands and coastal mainlands[57], causing thousands of species extinctions and altering trophic structures[58]. On continents, invasive species currently have a greater effect on biodiversity in economically developed and extra-tropical regions, but tropical ecosystems are predicted to become increasingly vulnerable to invasion as the 21st century continues[59]. Despite a deficit of research in the tropics[60], two prominent examples highlight the scope and magnitude of species invasions into terrestrial tropical ecosystems: there was an 84% increase in detections of alien species between 2003 and 2010 in Singapore[61], and invasive African grasses could threaten up to 380,000 km^2 of Australia's savannahs by promoting landscape flammability[62]. In aquatic ecosystems, invasive predatory fish—such as the Indo-Pacific lionfish in Caribbean coral reefs[63] or the Nile perch in African lakes[64]—have contributed to the loss of native species. Marine invasions are also facilitated by the mass transport of species in the ballast water of ships, which results in widespread biotic homogenization[65].

Global climate change

Many of these local stressors are promoted by globalized drivers, but climate change is truly global. Increases in atmospheric CO_2 concentrations to levels higher than 400 p.p.m. have important implications for tropical terrestrial and aquatic ecosystems. Ocean acidification from dissolved CO_2 is changing ocean chemistry to the extent that declining coral calcification has already been detected[66]. Conditions for reef accretion and growth may be mostly absent throughout the tropics by the year 2100 under 'business-as-usual' emission scenarios[67]. Within savannahs, elevated CO_2 levels favour the growth of woody plants over grasses, contributing to woody encroachment and the potential for a switch in biome state[68,69]. CO_2 fertilization may also have contributed to enhanced tree productivity and mortality rates observed in humid tropical forests[70].

Global warming is not proceeding at the same rate across the planet. Although the greatest absolute temperature increases are occurring

© 2018 Springer Nature Limited. All rights reserved.

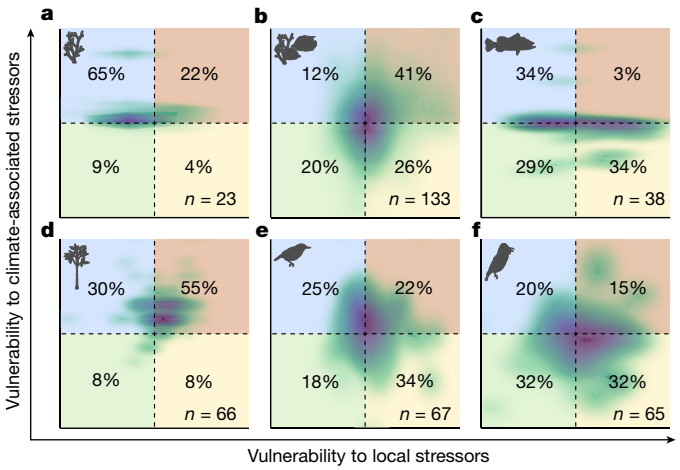


Fig. 3 | Vulnerability of tropical biota to local and climatic stressors. Species co-tolerance to a local and climate-associated stressor[85]. The *x* axis shows responses to fishing for corals (**a**), reef fish (**b**) and freshwater fish (**c**); changes in landscape configuration for small-stemmed trees (diameter at breast height between 2 and 10 cm (**d**)) and forest birds (**e**); and fire suppression for savannah birds (**f**). The *y* axis represents longitudinal responses to climate-associated events: the 2015–2016 and 1997–1998 coral bleaching events in the Seychelles for corals (**a**) and reef fish (**b**), respectively; the 1997–1998 El Niño-induced drought for lower Amazonian freshwater fish (**c**); Amazonian fires during the 2015–2016 El Niño event for small-stemmed trees (**d**) and forest birds (**e**); and shrub encroachment between 1998–2008 in South Africa for savannah birds (**f**). Species relative density is represented from low (light green) to high (dark blue). The four quadrants represent the location of 'survivor' species tolerant to both stressors (green), species only susceptible to local stressors (yellow), species only vulnerable to climate-associated stressors (blue) and 'double jeopardy' species susceptible to both stressors (red). Numbers show the percentage of species that fall into the quadrant. *n* gives the total number of species (genera for corals). Data sources are presented in Supplementary Table 1.

at higher latitudes, the tropics are already some of the hottest places on the planet and have the lowest inter-annual temperature variability[71,72]. Consequently, they will be the first areas to experience significantly warmer climates than the present day[72] and will endure climatic conditions without present-day equivalents[71]. In addition, some of the most important climate oscillations—including El Niño and the Indian Ocean Dipole—take place within and exert their greatest influence on tropical regions. It is unclear whether these oscillations will change in a warming world, but extremes of their phases have the potential to exacerbate or ameliorate the overall warming trend. One outcome of increasing temperatures is the pole-ward shift of species ranges or movement to higher altitudes or deeper depths[73]. For example, corals in southern Japan are extending northwards at about 14 kilometres per year[74], and temperate macroalgal communities are being replaced with corals and other tropical species along large stretches of Australian coastline[75]. Latitudinal shifts in terrestrial and freshwater tropical species distributions are less certain, because of the many natural and anthropogenic barriers to species movement and the low dispersal capacity of many tropical species[76]. Furthermore, the responses of terrestrial species are defined by changes in rainfall as well as temperature[77].

If movement is not an option, tropical species must adapt or face extinction. Unfortunately, there is evidence that some species are either approaching their physiological limits or are unable to adapt to the current rate of environmental change[78]. Increasing ocean temperature extremes are driving mass-bleaching events and mortality of reef-forming corals, with the time between bleaching events declining by 76–80% since the early 1980s[79]. Higher temperatures also affect tropical vertebrates—for example, by causing an extreme female bias in the sex ratio of green turtles in the warmer regions of the Great Barrier Reef[80] and a reduction in the reproductive success of African wild dogs[81]. Altered rainfall is also critical. Droughts are drying up biologically diverse small streams[82], and even modest changes in dry-season length increase tropical tree mortality[70] and modify tropical forest bird community structure[83].

Interacting stressors and indirect effects

Stressors affecting tropical species can interact in myriad ways[84]. We demonstrate this by compiling data from six case studies within a co-tolerance framework that enables the examination of species responses to two dominant stressors[85]. Only a small subset of species or genera (8–32%) showed no or positive responses when both stressors were combined (Fig. 3), and up to 55% fell within the 'double jeopardy' quadrant that indicates a negative response to both stressors. Although our summary does not quantify the magnitude of effects, it clearly demonstrates that stressors can act together to reduce the abundance or occupancy of tropical species. Moreover, most tropical ecosystems are affected by more than two stressors at any given location and time[84], and co-tolerance analyses of this type are likely to underestimate the reality of human impacts.

Many changes to tropical ecosystems result from the indirect consequences of single or multiple stressors. On coral reefs, nutrient inputs from the land may increase susceptibility to coral bleaching, disease and outbreaks of pests[86], and pole-ward reef expansion is supported by feedbacks from range-shifts in tropical herbivorous fish[75]. Overexploitation can result in surprising changes in tropical ecosystem properties through trophic cascades. For instance, the extirpation of a single detritivore fish species in the Orinoco basin reduced downstream organic-carbon transport, which in turn increased net primary productivity and respiration[87]. On reefs, the overfishing of keystone predators has repercussions for benthic structure[88], and the removal of herbivores can limit coral recovery from mass-mortality events[89]. In mesic savannahs, changes to herbivore numbers alter ecosystem functions and structure through their interactions with wildfire regimes[90]. Invasive species are also frequently linked to other stressors: the introduction of the Nile perch had a major role in the decline of endemic fish species in Lake Victoria, but these declines were also catalysed by a combination of other drivers including soil erosion, eutrophication and overfishing[64].

Ecosystems in transition

Interactions between multiple anthropogenic stressors are causing pervasive changes in the tropics, such that alternative states are emerging across all major tropical ecosystems (Fig. 4). Perhaps counter-intuitively, trees are encroaching on savannahs and grasses are invading disturbed tropical forests—however, in both cases the changes are from species-rich to species-poor systems[68,91].

These marked ecosystem transitions are accompanied by widespread modification of species composition. For example, the relative abundance of coral species has been altered on reefs that maintain coral dominance[92]; the extirpation of native fish has followed species introductions in lakes[64]; liana biomass has increased in otherwise undisturbed Neotropical forests[93]; and patterns of plant regeneration in humid forests have been altered by the overharvesting of seed-dispersing vertebrates[31,94]. Altered species composition is a cause for concern because it could signal the onset of more severe modification, especially if dominant species are vulnerable to other stressors or if there are cascading implications for ecosystem functioning. The collapse of Jamaican coral reefs provides one of the starkest examples. First, chronic overfishing depleted herbivorous fish populations, which left the system over-reliant on sea urchins for grazing algae. Then, in 1980, Hurricane Allen affected the system, creating a substantial amount of dead substrate. Although corals began to recover after the hurricane, the subsequent mass mortality of sea urchins owing to disease—combined with the fact that herbivorous fish were already at low levels of abundance—led to a phase shift from coral to macroalgal dominance[95,96].

© 2018 Springer Nature Limited. All rights reserved.

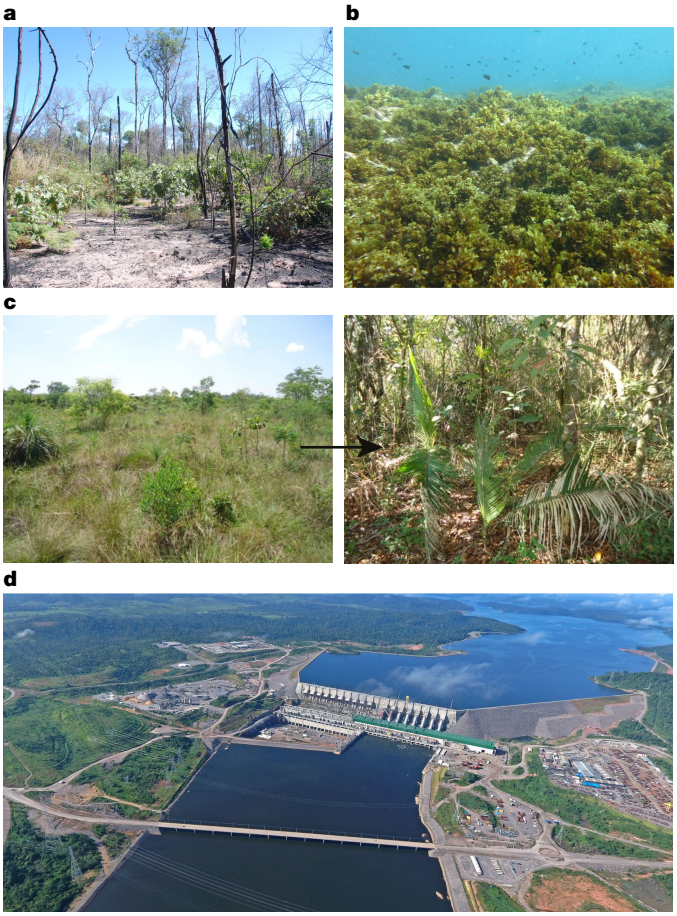


Fig. 4 | Tropical ecosystems in transition. a, Recurrent wildfires in historically fire-free humid tropical forests[167] can lead to the dominance of grassy vegetation that impedes succession towards closed-canopy forests[91,168]. These wildfires result from the combination of local actions (for example, agricultural practices and logging) and climate change that has increased the prevalence of weather that promotes wildfires[169]. **b**, Chronic local pressures and acute climatic stressors can lead to coral cover being replaced by macroalgae, sponges or sediment-laden turf algae[89,95]. During the 1998 global coral-bleaching event, >90% of live coral died in the inner Seychelles and nearly half of the reefs transitioned to fleshy macroalgal regimes[89]. **c**, Woody encroachment is occurring in many savannahs[69], causing biodiversity loss and altered system functioning[68]. Causes are mixed: regime shifts to forest-associated ecosystems have been attributed to fire suppression policies (for example, Brazilian cerrado (left) to forest (right)[170]), or changes in herbivory and increasing atmospheric CO_2[69]. **d**, The boom in hydropower dam construction is affecting large tropical river basins[134]. The transformation from lotic to lentic conditions reduces access to riparian and floodplain habitats that are nursery areas and feeding grounds for many of the species occupying higher trophic levels, leading to major shifts in species composition and ecosystem function[82]. Images from J.B. (**a**), N.A.J.G. (**b**), G. Durigan (**c**) and C.G.L. (**d**), used with permission.

Socio–economic context and response capacity

The interacting proximate stressors causing tropical environmental change are underpinned by broader changes in socio-economic and political factors. We examined the trajectories of four types of underlying distal drivers, including demography and the economy (Fig. 5a, b), socio-political factors (Fig. 5c, d), markets (Fig. 5e, f) and technology[97] (Fig. 5g, h) to explore how tropical countries are changing relative to the rest of the world and to evaluate the relative influence of local and global drivers. We also examined how the capacity of tropical countries to reduce or cope with proximate stressors compares to non-tropical countries based on underlying governance systems (Fig. 5i, j) and research capacity (Fig. 5k, l).

The immense biodiversity of the tropics exists in the context of rapid demographic and economic growth (Fig. 5a, b). The human population is growing at a faster rate in the tropics than elsewhere (Fig. 5a), and by 2050 half of the world's population will live in the tropics[2]. These demographic changes are accompanied by a steady growth in gross domestic product (GDP) that is linked in part to the rapid expansion of agricultural and extractive industries. However, in the tropics, per capita GDP—which is an important measure of human well-being—remains far lower than the non-tropical average (Fig. 5b) and the rates of change suggest that there has been little closing of the global inequality gap. Although the relationship between development and natural resource conservation does not have to be negative[98,99], measures that reflect higher social performance are almost always associated with higher resource use[99]. A larger and more affluent tropical population will increase demands for timber, water, food, energy and land, all of which are strongly linked with environmental degradation.

These internal changes will be exacerbated by economic growth in non-tropical countries and the continued displacement of environmental effects to less-developed areas[100]. Indeed, despite high levels of tropical cultural diversity[21,22], external socio-political influences (Fig. 5c, d) suggest that tropical countries have become increasingly susceptible to globalization. For example, the proportion of imported food crops (Fig. 5c) and foreign land acquisitions are far higher in the tropics than elsewhere (Fig. 5d) and are associated with extensive road building[101] and agricultural investment[102]. These trends towards increasing tropical globalization are reinforced by changes in market integration (Fig. 5e, f) and technological development (Fig. 5g, h). For example, agricultural exports (Fig. 5f) are steadily increasing, albeit from a far lower baseline than the rest of the world. Moreover, given comparatively low levels of adoption of technological developments, such as industrial fishing techniques (Fig. 5g) or fertilizers (Fig. 5h), there is an enormous risk that the rate of natural resource extraction in many tropical countries will increase further, to supply both domestic and export markets[103,104]. Taken together, these examples highlight the crucial role that external markets will have in determining the fate of tropical ecosystems.

Effective environmental governance (Fig. 5i, j) is a necessary condition for improved sustainability outcomes[105], particularly when domestic (Fig. 5a, b) and global (Fig. 5c–f) distal drivers are expected to exert increasing and unsustainable pressure on tropical ecosystems[2,102]. However, national-level assessments of governance effectiveness place the tropics far below extra-tropical countries, with no sign of improvement (Fig. 5i). External support for environmental governance may help where local governance is weak. Yet, despite the biological importance of the tropics, levels of environmental aid from the Organisation for Economic Cooperation and Development (OECD) are only marginally greater in the tropics than elsewhere (Fig. 5j), and these investments are dwarfed by the value of domestic resource extraction (for example, agricultural exports; Fig. 5f), the value of which is two orders of magnitude greater than overseas environmental aid. Furthermore, OECD environmental aid has been declining in recent years and seems unlikely to increase in the short term[106].

Low governance capacity in the tropics is further exacerbated by insufficient research and development investment (Fig. 5k) and low levels of scientific output (Fig. 5l). Research investment is critical for driving innovation and the development of evidence-based solutions to environmental degradation[107]. Despite some notable centres of excellence, the vast majority of biodiversity-related data and research is concentrated in wealthy, non-tropical countries[17], and manuscripts submitted by authors from low-income countries are less than half as likely to be published as those from high-income countries[108]. These trends highlight an alarming disconnect between the global scientific process and the people that are most capable of engaging with decision makers in tropical countries, who have the best understanding of local context and, arguably, have the strongest incentive to achieve positive outcomes for tropical conservation through their research.

© 2018 Springer Nature Limited. All rights reserved.

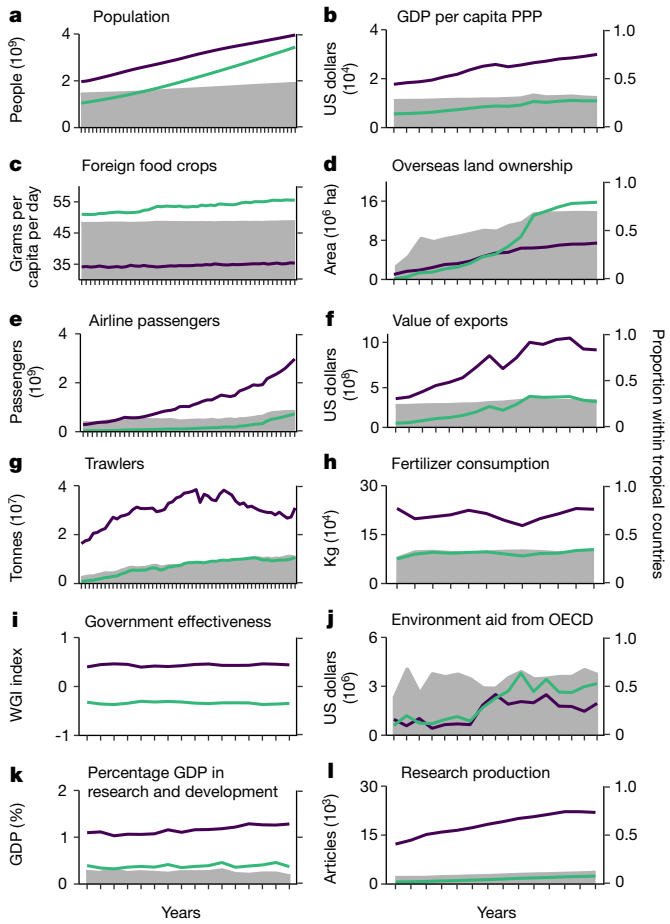


Fig. 5 | Socio-economic drivers of biodiversity loss and societal response capacities. Green lines represent countries with >50% of their area within tropical latitudes; purple lines represent all other countries; grey-shaded areas represent the proportion of the global total within tropical countries. **a**, Population (1960–2016). **b**, GDP per capita (2011 US dollars, based on purchasing power parity; 2000–2016). **c**, Foreign food crops (1961–2009). **d**, Cumulative overseas land ownership (2001–2017). **e**, Domestic and international airline passengers (1970–2016). **f**, Agricultural and forestry commodities export value (2001–2016). **g**, Bottom and pelagic trawler catch tonnages (1960–2014). **h**, Total fertilizer (nitrogen, potash and phosphate) consumption relative to crop area (2002–2013). **i**, Government effectiveness index (2000–2016). WGI, World Governance Indicators. **j**, Environmental aid (2000–2016). **k**, Public and private sector research and development expenditure as a percentage of GDP (2000–2015). **l**, Scientific and technical journal articles per million people in the fields of physics, biology, chemistry, mathematics, clinical medicine, biomedical research, engineering and technology, and Earth and space sciences (2003–2016). Data sources are presented in Supplementary Table 1.

Diverse solutions for diverse systems

Tropical ecosystems—and therefore at least 78% of global biodiversity (Fig. 2a)—are at a critical juncture. Multiple interacting local and global stressors (Fig. 3) that are driving species extinctions and potentially irreversible ecosystem transitions[92,109] (Fig. 4) are set within a changing socio-economic context (Fig. 5). This changing context is characterized by growing and more affluent populations, an increasingly globalized world, and weak governance and research capacity— all of which threaten to increase environmental degradation, conflict and inequality[102]. Countering these threats requires major improvements in local and global governance capacity and a step-change in how environmental objectives are integrated into broader development goals[110]. We review the opportunities and limitations presented by three well-established and non-mutually exclusive approaches to conservation, before highlighting priorities for research.

Conservation approaches

A fundamental element of tropical conservation relies on protected areas to limit demographic pressures and the effect of local stressors. These are supported by a wealth of scientific evidence outlining the pervasive effect of local stressors across tropical ecosystems[37,49] (Fig. 3) combined with an eco-centric philosophy that emphasizes the intrinsic rights of nature[111]. Despite a substantial expansion of protected area coverage in the marine and forested tropics[112], the current network remains poorly designed, has very limited coverage of tropical freshwaters and grasslands, and is inadequately resourced[113]. Moreover, a strategy focused solely on protected areas may not foster environmental conservation outside of reserves[114] and fails to engage with the distal drivers of biodiversity loss (Fig. 5) that can undermine the effectiveness of protected areas themselves[115].

A second set of approaches for tropical conservation is based on the notion that people need to perceive the benefits of nature to justify conservation. These approaches emphasize the need to pursue conservation objectives in human-dominated landscapes, the provision of ecosystem services and the involvement of private-sector actors. In the tropics, they are epitomized by the growth in market-based conservation payment mechanisms, such as REDD+[116], investments in the 'blue economy'[117] and a step-change in the number of companies making sustainability commitments[118]. These approaches have strengthened the conservation toolkit, especially where strict regulatory approaches have failed. Encouraging examples range from the positive effects of commodity certification (for example, palm oil[119]) to payment for ecosystem service schemes (for example, watershed protection[120]). However, such approaches also attract considerable criticism, with implementation often lagging behind commitments[118], persistent concerns around the social legitimacy of compensation schemes[121] and the misalignment of market-based mechanisms with local needs and perceptions of environmental values[122].

A third and more diverse set of approaches is based on the recognition of the interdependencies between people and nature, the coevolution of ecological and socio-economic systems at local, regional and global scales[123], and perspectives about the co-existence of people and nature. This set of 'systems-based' approaches includes: (1) an appreciation of the importance of bottom-up, community-based conservation approaches in human-dominated land- and seascapes (for example, small-scale fisheries[124] and community-managed forests[125]); (2) recognition of the role of indigenous people as environmental stewards and shifts towards an appreciation of more collective relationships with nature (for example, the Ecuadorian constitution[126]); (3) landscape- and ecosystem-wide approaches that attempt to bridge the role of actors working at different scales and in different sectors (for example, jurisdictional approaches to curb deforestation[127]); and (4) a more explicit accounting of multi-scale feedbacks, including the role of distant market actors and distal drivers[123]. These broad, multi-layered 'people and nature' approaches hold considerable appeal, but the inherent complexity of local contexts can make them challenging to conceptualize, implement and measure in joined-up and consistent ways[128].

Acting together and acting now

The three approaches to the conservation and governance of tropical ecosystems outlined above are often associated with alternative researcher and practitioner worldviews[129,130]. However, the ecological diversity (Fig. 2a), vulnerability (Figs. 2b, 3) and socio-economic complexity (Fig. 5) of the tropics highlights the importance of pluralism[131] and the need to adopt a variety of what are often complementary and synergistic approaches[130]. For all their deficiencies, protected areas are indispensable to limit the effect of local stressors, and it will be impossible to avoid further biodiversity loss unless they are strengthened and expanded[132]. However, conservation strategies must also address the underlying drivers of environmental change (Fig. 5) and avoid

© 2018 Springer Nature Limited. All rights reserved.

exacerbating deeply rooted inequalities[114]. Practice is always messier than theory, and the adoption of more sustainable management systems is usually only possible with the support of a range of actors, as can be seen in the recent successes of some hybrid governance approaches, with government, the private sector and civil society organizations all having vital roles[133].

Another clear message is that conservation efforts need to operate at local, regional and global scales to be effective. Many distal drivers are disconnected in both space and time from the sites they affect, and the engagement of external actors—including in distant markets and governance processes—is often essential to ensure that local efforts are effective. These include more strategic integration of environmental policy with development goals[134], the need for multinational environmental governance approaches, especially for aquatic systems[82], and recognition of the importance of tackling demand for unsustainable products from downstream buyers and investors[118]. The capstone of such efforts lies in the urgent need to deliver on the Paris Agreement, without which climate change will undercut or even negate hard-won local conservation successes, whether in coral reefs[92] or tropical forests[109].

Finally, we need to act now to address the pressing environmental challenges facing the tropics. This means being adaptive, learning by doing and embracing innovation. The past decades have seen a boom in proposals, innovations and insights about the governance and management of tropical ecosystems, ranging from more technocentric proposals to facilitate the evolution of climate-tolerant corals[135]; ecological engineering to recover lost trophic interactions by species re-introductions, ecological replacements and rewilding[136]; to radical new legal frameworks such as France's 'Loi de vigilance' (2017-399; https://www.legifrance.gouv.fr/eli/loi/2017/3/27/2017-399/jo/texte) that places an unprecedented due diligence obligation on major companies to assess social and environmental risks in their supply chains that extend beyond French borders. Though these innovations serve different purposes and are varyingly scalable, they illustrate the potential of solutions-based science and conservation. Of course, acting now does not mean ignoring the existing evidence base or making uninformed decisions. Rather, it is vital that researchers and decision makers are vigilant to opportunities and risks and are willing to learn lessons.

Keeping pace with the Anthropocene

All approaches to governing tropical ecosystems will be more effective if they have local support and are based on strong scientific evidence that ensures, for example, that protected areas are located where they are most needed, ecosystem services are accurately quantified, extractive activities such as fishing and logging are managed sustainably, and underlying drivers of environmental degradation are identified and understood. Although these challenges are common to all conservation and sustainability science, they are magnified in the tropics owing to the unique diversity and high vulnerability of tropical ecosystems and the low research capacity of most tropical countries. Here we examine four areas in which research effort can be more closely aligned with some of the priorities highlighted by this review.

Addressing key knowledge shortfalls

Our understanding of tropical biodiversity is limited by substantial shortfalls in knowledge regarding taxonomy and species distributions[137]. Overcoming these shortfalls will require targeting resources towards the data-scarce regions that cover so much of the tropics[18]. At the ecosystem level, there is a need for increased study of structurally and functionally distinct systems, particularly tropical grassy biomes[68], dry forests[138] and low-order stream systems[139]. Progress in these areas is likely to be aided by advances in DNA sequencing and informatics, which have the potential to invigorate taxonomic discovery, and by reaching across cultural divides to incorporate national, regional and local knowledge that often remains ignored because it is not available in English[140], included in standard databases[141] or recognized by conventional science[142].

Understanding vulnerability

Our growing knowledge of the role of individual stressors, such as landscape configuration or overexploitation, needs to be complemented by research on the effect of multiple stressors[84], which could help predict and mitigate complex biotic responses when climate and local stressors act in concert (Fig. 3). Other phenomena that are important but harder to study include the role of time lags or extinction debts[40], trophic cascades[31] and trajectories of ecosystem degradation and recovery in the face of unprecedented environmental change[143]. Revealing these more-complex forms of vulnerability will often demand longer-term and larger multi-scale sampling and monitoring programs. New approaches are also needed to overcome one of the more intractable challenges of tropical ecology: the fact that we often know least about the rarest and most vulnerable species or taxonomic groups.

Understanding distal drivers

Conservation does not occur in a vacuum, and local interventions are likely to be much more effective if they are guided by a closer understanding of underlying distal drivers of biodiversity loss and environmental change, including identifying the actors behind such drivers, which will help to determine potential trigger points and identify more effective policy responses[97]. Unpicking the role of distal drivers is essential to understand how interactions between social and environmental systems shape local environmental outcomes[144]. Careful study has revealed many surprising interactions, such as links between the intensification of commercial fishing and increased bushmeat exploitation in west Africa[145], the role of warfare in driving African mammal declines[146] or the role of currency exchange rates in driving deforestation[147]. Achieving this deeper understanding requires greater integration of the natural and social sciences, with interdisciplinarity included as a core element of tropical conservation research[148].

From research to impact

Achieving positive effects from conservation research relies on building a stronger interface between science and society that challenges the oversimplified assumption of a linear flow from knowledge to action[149]. Engendering positive changes will require closer participation of practitioners in the research process and investments in outreach activities and professional capacity building[149]. These will be supported by studying the knowledge exchange process itself, including the critical part that is played by knowledge brokers and boundary organizations[150–152]. Part of this process will require a focus on success stories or 'bright spots', which will enable the social, institutional and environmental conditions that create positive outcomes to be identified and replicated[151]. The positive social and ecological outcomes from innovative restoration and rewilding programmes in Costa Rica and Mozambique demonstrate the potential for positive action[153].

Local managers and scientists have a vital role in designing and implementing research that can inform regionally appropriate conservation actions[154]. At present, our knowledge of hyperdiverse ecosystems is over-reliant on inferences gleaned from distant research stations or inappropriate theoretical constructs developed for temperate contexts[18,155]. Research is also more likely to have an effect if the spatial scale of studies is closely matched to the administrative scale at which resource decisions are taken[156]. Sustaining research programmes and learning networks in study landscapes can help build vital relationships between researchers, local knowledge holders and decision makers[154].

Achieving these changes requires building on trends in the technological, disciplinary and cultural dimensions of research practice. In the technological domain, opportunities for data collection have been revolutionized by developments in remote sensing and drones[157], the plummeting costs of DNA technologies[158] and the step-changes in bioinformatics that have enabled 'big data' to be stored and retrieved in open-access platforms[159]. In the disciplinary domain, the past decade has seen a marked upward trend in interdisciplinary and transdisciplinary research and a greater—albeit still insufficient—integration of

© 2018 Springer Nature Limited. All rights reserved.

natural and social sciences. This has resulted in an increasing openness of researchers towards methodological pluralism and mixed-method approaches[149] and a growing recognition of the contribution that can be made by local people and citizen- and para-scientists in biodiversity research[160]. Changes in research culture include the greater internationalization of ecological science and closer approximation with society[149], both of which can help foster more fertile ground for knowledge exchange and capacity building. Notable advances include the development of multi-disciplinary and multinational learning networks[161], exponential growth in author teams[162] and major syntheses such as the Intergovernmental Platform for Biodiversity and Ecosystem Services.

Recent years have seen an awakening of environmental consciousness and calls for decisive action that are manifest, for example, in the Paris Agreement, the Sustainable Development Goals and voluntary Zero Deforestation Commitments. Scientists from tropical and non-tropical regions can inform these endeavours by developing a reliable knowledge base and innovative management interventions. Overcoming the remaining research challenges is far from trivial and will require a massive investment of resources to develop scientific infrastructure and capacity within tropical nations, as well as profound changes to ways of working and the relationship between the research process and society at large. But a failure to act decisively and to act now will greatly increase the risk of unprecedented and irrevocable biodiversity loss in the hyperdiverse tropics.

Received: 9 March 2018; Accepted: 24 May 2018;
Published online 25 July 2018.

1. United Nations General Assembly. *Transforming our world: the 2030 Agenda for Sustainable Development* (United Nations, New York, 2015).
2. Edelman, A. et al. *State of the Tropics: 2014 Report* (James Cook Univ., Cairns, 2014).
3. Moreau, C. S. & Bell, C. D. Testing the museum versus cradle tropical biological diversity hypothesis: phylogeny, diversification, and ancestral biogeographic range evolution of the ants. *Evolution* **67**, 2240–2257 (2013).
4. Jablonski, D. et al. Out of the tropics: evolutionary dynamics of the latitudinal diversity gradient. *Science* **314**, 102–106 (2006).
5. DeGraaf, R. M. & Rappole, J. H. *Neotropical Migratory Birds: Natural History, Distribution, and Population Change* (Cornell Univ. Press, Ithaca, 1995).
6. Hahn, S., Bauer, S. & Liechti, F. The natural link between Europe and Africa – 2.1 billion birds on migration. *Oikos* **118**, 624–626 (2009).
7. Tittensor, D. P. et al. Global patterns and predictors of marine biodiversity across taxa. *Nature* **466**, 1098–1101 (2010).
8. Chown, S. L., Gaston, K. J. & Williams, P. H. Global patterns in species richness of pelagic seabirds: the Procellariiformes. *Ecography* **21**, 342–350 (1998).
9. Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B. & Worm, B. How many species are there on Earth and in the ocean? *PLoS Biol.* **9**, e1001127 (2011).
This study develops a new method to quantify the completeness of taxonomic inventories.
10. Tancoigne, E. & Dubois, A. Taxonomy: no decline, but inertia. *Cladistics* **29**, 567–570 (2013).
11. Jenkins, C. N., Pimm, S. L. & Joppa, L. N. Global patterns of terrestrial vertebrate diversity and conservation. *Proc. Natl Acad. Sci. USA* **110**, E2602–E2610 (2013).
12. Guénard, B., Weiser, M. D. & Dunn, R. R. Global models of ant diversity suggest regions where new discoveries are most likely are under disproportionate deforestation threat. *Proc. Natl Acad. Sci. USA* **109**, 7368–7373 (2012).
This paper identifies hotspots of discovery for ant species.
13. Burgin, C. J., Colella, J. P., Kahn, P. L. & Upham, N. S. How many species of mammals are there? *J. Mamm.* **99**, 1–14 (2018).
This paper shows that the Neotropics and Afrotropics contain the highest number of newly recognized mammal species.
14. Fisher, R. et al. Species richness on coral reefs and the pursuit of convergent global estimates. *Curr. Biol.* **25**, 500–505 (2015).
15. Agnarsson, I., Coddington, J. A. & Kuntner, M. in *Spider Research in the 21st Century: Trends and Perspectives* (ed. Penney, D.) 58–111 (Siri Scientific, Rochdale, 2013).
16. Collen, B., Ram, M., Zamin, T. & McRae, L. The tropical biodiversity data gap: addressing disparity in global monitoring. *Trop. Conserv. Sci.* **1**, 75–88 (2008).
17. Fisher, R. et al. Global mismatch between research effort and conservation needs of tropical coral reefs. *Conserv. Lett.* **4**, 64–72 (2011).
18. Gardner, T. A. et al. Prospects for tropical forest biodiversity in a human-modified world. *Ecol. Lett.* **12**, 561–582 (2009).
19. Barlow, J., Overal, W. L., Araujo, I. S., Gardner, T. A. & Peres, C. A. The value of primary, secondary and plantation forests for fruit-feeding butterflies in the Brazilian Amazon. *J. Appl. Ecol.* **44**, 1001–1012 (2007).
20. Baldwin, C. C., Tornabene, L. & Robertson, D. R. Below the mesophotic. *Sci. Rep.* **8**, 4920 (2018).
21. Collard, I. F. & Foley, R. A. Latitudinal patterns and environmental determinants of recent human cultural diversity: do humans follow biogeographical rules? *Evol. Ecol. Res.* **4**, 371–383 (2002).
22. Simons, G. F. & Fennig, C. D. *Ethnologue: Languages of the World* 20th edn (SIL International, Dallas, 2018).
23. Purugganan, M. D. & Fuller, D. Q. The nature of selection during plant domestication. *Nature* **457**, 843–848 (2009).
24. Burke, L., Reytar, K., Spalding, M. & Perry, A. *Reefs at Risk Revisited* (WRI, Washington, D.C., 2011).
25. Ferrario, F. et al. The effectiveness of coral reefs for coastal hazard risk reduction and adaptation. *Nat. Commun.* **5**, 3794 (2014).
26. Bonan, G. B. Forests and climate change: forcings, feedbacks, and the climate benefits of forests. *Science* **320**, 1444–1449 (2008).
27. Grace, J., San José, J., Meir, P., Miranda, H. S. & Montes, R. A. Productivity and carbon fluxes of tropical savannas. *J. Biogeogr.* **33**, 387–400 (2006).
28. van der Ent, R. J., Savenije, H. H. G. G., Schaefli, B. & Steele-Dunne, S. C. Origin and fate of atmospheric moisture over continents. *Wat. Resour. Res.* **46**, W09525 (2010).
This study shows the importance of tropical forests for precipitation in extra-tropical regions.
29. IUCN. IUCN Spatial Data. http://www.iucnredlist.org/technical-documents/spatial-data (2018).
30. Régnier, C. et al. Mass extinction in poorly known taxa. *Proc. Natl Acad. Sci. USA* **112**, 7761–7766 (2015).
31. Dirzo, R. et al. Defaunation in the Anthropocene. *Science* **401**, 401–406 (2014).
32. Keenan, R. J. et al. Dynamics of global forest area: results from the FAO Global Forest Resources Assessment 2015. *For. Ecol. Manage.* **352**, 9–20 (2015).
33. Steffen, W., Broadgate, W., Deutsch, L., Gaffney, O. & Ludwig, C. The trajectory of the Anthropocene: the great acceleration. *Anthropocene Rev.* **2**, 81–98 (2015).
34. Pelicice, F. M., Pompeu, P. S. & Agostinho, A. A. Large reservoirs as ecological barriers to downstream movements of Neotropical migratory fish. *Fish Fish.* **16**, 697–715 (2015).
35. Mendenhall, C. D., Shields-Estrada, A., Krishnaswami, A. J. & Daily, G. C. Quantifying and sustaining biodiversity in tropical agricultural landscapes. *Proc. Natl Acad. Sci. USA* **113**, 14544–14551 (2016).
36. Pfeifer, M. et al. Creation of forest edges has a global impact on forest vertebrates. *Nature* **551**, 187–191 (2017).
37. Barlow, J. et al. Anthropogenic disturbance in tropical forests can double biodiversity loss from deforestation. *Nature* **535**, 144–147 (2016).
38. Spracklen, D. V. & Garcia-Carreras, L. The impact of Amazonian deforestation on Amazon basin rainfall. *Geophys. Res. Lett.* **42**, 9546–9552 (2015).
39. Haddad, N. M. et al. Habitat fragmentation and its lasting impact on Earth's ecosystems. *Sci. Adv.* **1**, e1500052 (2015).
40. Gibson, L. et al. Near-complete extinction of native small mammal fauna 25 years after forest fragmentation. *Science* **341**, 1508–1510 (2013).
41. Dudgeon, D. et al. Freshwater biodiversity: importance, threats, status and conservation challenges. *Biol. Rev. Camb. Philos. Soc.* **81**, 163–182 (2006).
42. Altieri, A. H. et al. Tropical dead zones and mass mortalities on coral reefs. *Proc. Natl Acad. Sci. USA* **114**, 3660–3665 (2017).
43. Lewis, S. E., Silburn, D. M., Kookana, R. S. & Shaw, M. Pesticide behavior, fate, and effects in the tropics: an overview of the current state of knowledge. *J. Agric. Food Chem.* **64**, 3917–3924 (2016).
44. Bebber, D. P., Holmes, T. & Gurr, S. J. The global spread of crop pests and pathogens. *Glob. Ecol. Biogeogr.* **23**, 1398–1407 (2014).
45. Lebreton, L. C. M. et al. River plastic emissions to the world's oceans. *Nat. Commun.* **8**, 15611 (2017).
46. Hall, N. M., Berry, K. L. E., Rintoul, L. & Hoogenboom, M. O. Microplastic ingestion by scleractinian corals. *Mar. Biol.* **162**, 725–732 (2015).
47. Lamb, J. B. et al. Plastic waste associated with disease on coral reefs. *Science* **359**, 460–462 (2018).
48. Garcia, L. C., Ribeiro, D. B., de Oliveira Roque, F., Ochoa-Quintero, J. M. & Laurance, W. F. Brazil's worst mining disaster: corporations must be compelled to pay the actual environmental costs. *Ecol. Appl.* **27**, 5–9 (2017).
49. MacNeil, M. A. et al. Recovery potential of the world's coral reef fishes. *Nature* **520**, 341–344 (2015).
50. Castello, L. et al. The vulnerability of Amazon freshwater ecosystems. *Conserv. Lett.* **6**, 217–229 (2013).
51. Ripple, W. J. et al. Collapse of the world's largest herbivores. *Sci. Adv.* **1**, e1400103 (2015).
52. Ripple, W. J. et al. Status and ecological effects of the world's largest carnivores. *Science* **343**, 1241484 (2014).
53. Ingram, D. J. et al. Assessing Africa-wide pangolin exploitation by scaling local data. *Conserv. Lett.* **11**, e12389 (2017).
54. Eaton, J. A. et al. Trade-driven extinctions and near-extinctions of avian taxa in Sundaic Indonesia. *Forktail* **31**, 1–12 (2015).
55. Barrett, M. A., Brown, J. L., Morikawa, M. K., Labat, J.-N. & Yoder, A. D. CITES designation for endangered rosewood in Madagascar. *Science* **328**, 1109–1110 (2010).
56. Bellard, C., Cassey, P. & Blackburn, T. M. Alien species as a driver of recent extinctions. *Biol. Lett.* **12**, 20150623 (2016).
57. Dawson, W. et al. Global hotspots and correlates of alien species richness across taxonomic groups. *Nat. Ecol. Evol.* **1**, 0186 (2017).
58. Simberloff, D. et al. Impacts of biological invasions: what's what and the way forward. *Trends Ecol. Evol.* **28**, 58–66 (2013).
59. Early, R. et al. Global threats from invasive alien species in the twenty-first century and national response capacities. *Nat. Commun.* **7**, 12485 (2016).

© 2018 Springer Nature Limited. All rights reserved.

60. Pyšek, P. et al. Geographical and taxonomic biases in invasion ecology. *Trends Ecol. Evol.* **23**, 237–244 (2008).
61. Nghiem, T. P. et al. Economic and environmental impacts of harmful non-indigenous species in southeast Asia. *PLoS ONE* **8**, e71255 (2013).
62. Setterfield, S. A., Rossiter-Rachor, N. A., Hutley, L. B., Douglas, M. M. & Williams, R. J. Turning up the heat: the impacts of *Andropogon gayanus* (gamba grass) invasion on fire behaviour in northern Australian savannas. *Divers. Distrib.* **16**, 854–861 (2010).
63. Albins, M. A. & Hixon, M. A. Worst case scenario: potential long-term effects of invasive predatory lionfish (*Pterois volitans*) on Atlantic and Caribbean coral-reef communities. *Environ. Biol. Fishes* **96**, 1151–1157 (2013).
64. Hecky, R. E., Mugidde, R., Ramlal, P. S., Talbot, M. R. & Kling, G. W. Multiple stressors cause rapid ecosystem change in Lake Victoria. *Freshw. Biol.* **55**, 19–42 (2010).
65. Drake, J. M. & Lodge, D. M. Global hot spots of biological invasions: evaluating options for ballast-water management. *Proc. R. Soc. Lond. B* **271**, 575–580 (2004).
66. Albright, R. et al. Reversal of ocean acidification enhances net coral reef calcification. *Nature* **531**, 362–365 (2016).
67. Doney, S. C. et al. Climate change impacts on marine ecosystems. *Ann. Rev. Mar. Sci.* **4**, 11–37 (2012).
68. Parr, C. L., Lehmann, C. E. R., Bond, W. J., Hoffmann, W. A. & Andersen, A. N. Tropical grassy biomes: misunderstood, neglected, and under threat. *Trends Ecol. Evol.* **29**, 205–213 (2014).
This analysis highlights the importance of grassy tropical ecosystems.
69. Stevens, N., Lehmann, C. E. R., Murphy, B. P. & Durigan, G. Savanna woody encroachment is widespread across three continents. Glob. *Chang. Biol.* **23**, 235–244 (2017).
70. McDowell, N. et al. Drivers and mechanisms of tree mortality in moist tropical forests. *New Phytol.* (2018).
71. Williams, J. W., Jackson, S. T. & Kutzbach, J. E. Projected distributions of novel and disappearing climates by 2100 AD. *Proc. Natl Acad. Sci. USA* **104**, 5738–5742 (2007).
72. Mahlstein, I., Knutti, R., Solomon, S. & Portmann, R. W. Early onset of significant local warming in low latitude countries. *Environ. Res. Lett.* **6**, 034009 (2011).
73. Pecl, G. T. et al. Biodiversity redistribution under climate change: impacts on ecosystems and human well-being. *Science* **355**, eaai9214 (2017).
74. Yamano, H., Sugihara, K. & Nomura, K. Rapid poleward range expansion of tropical reef corals in response to rising sea surface temperatures. *Geophys. Res. Lett.* **38**, L04601 (2011).
75. Wernberg, T. et al. Climate-driven regime shift of a temperate marine ecosystem. *Science* **353**, 169–172 (2016).
76. Moore, R. P., Robinson, W. D., Lovette, I. J. & Robinson, T. R. Experimental evidence for extreme dispersal limitation in tropical forest birds. *Ecol. Lett.* **11**, 960–968 (2008).
77. VanDerWal, J. et al. Focus on poleward shifts in species' distribution underestimates the fingerprint of climate change. *Nat. Clim. Chang.* **3**, 239–243 (2013).
78. Hoffmann, A. A. & Sgrò, C. M. Climate change and evolutionary adaptation. *Nature* **470**, 479–485 (2011).
79. Hughes, T. P. et al. Spatial and temporal patterns of mass bleaching of corals in the Anthropocene. *Science* **359**, 80–83 (2018).
80. Jensen, M. P. et al. Environmental warming and feminization of one of the largest sea turtle populations in the world. *Curr. Biol.* **28**, 154–159 (2018).
81. Woodroffe, R., Groom, R. & McNutt, J. W. Hot dogs: high ambient temperatures impact reproductive success in a tropical carnivore. *J. Anim. Ecol.* **86**, 1329–1338 (2017).
82. Castello, L. & Macedo, M. N. Large-scale degradation of Amazonian freshwater ecosystems. Glob. *Chang. Biol.* **22**, 990–1007 (2016).
83. Brawn, J. D., Benson, T. J., Stager, M., Sly, N. D. & Tarwater, C. E. Impacts of changing rainfall regime on the demography of tropical birds. *Nat. Clim. Chang.* **7**, 133–136 (2017).
84. Côté, I. M., Darling, E. S. & Brown, C. J. Interactions among ecosystem stressors and their importance in conservation. *Proc. R. Soc. Lond. B* **283**, 20152592 (2016).
This paper outlines the importance of interactions between different stressors.
85. Vinebrooke, R. D. et al. Impacts of multiple stressors on biodiversity and ecosystem functioning: the role of species co-tolerance. *Oikos* **104**, 451–457 (2004).
86. Vega Thurber, R. L. et al. Chronic nutrient enrichment increases prevalence and severity of coral disease and bleaching. Glob. *Chang. Biol.* **20**, 544–554 (2014).
87. Taylor, B. W., Flecker, A. S. & Hall, R. O. Jr. Loss of a harvested fish species disrupts carbon flow in a diverse tropical river. *Science* **313**, 833–836 (2006).
88. McClanahan, T. R. et al. Critical thresholds and tangible targets for ecosystem-based management of coral reef fisheries. *Proc. Natl Acad. Sci. USA* **108**, 17230–17233 (2011).
89. Graham, N. A. J., Jennings, S., MacNeil, M. A., Mouillot, D. & Wilson, S. K. Predicting climate-driven regime shifts versus rebound potential in coral reefs. *Nature* **518**, 94–97 (2015).
90. Waldram, M. S., Bond, W. J. & Stock, W. D. Ecological engineering by a mega-grazer: white rhino impacts on a south african savanna. *Ecosystems* **11**, 101–112 (2008).
91. Veldman, J. W. & Putz, F. E. Grass-dominated vegetation, not species-diverse natural savanna, replaces degraded tropical forests on the southern edge of the Amazon Basin. *Biol. Conserv.* **144**, 1419–1429 (2011).
92. Hughes, T. P. et al. Coral reefs in the Anthropocene. *Nature* **546**, 82–90 (2017).
93. Phillips, O. L. et al. Increasing dominance of large lianas in Amazonian forests. *Nature* **418**, 770–774 (2002).
94. Harrison, R. D. et al. Consequences of defaunation for a tropical tree community. *Ecol. Lett.* **16**, 687–694 (2013).
95. Hughes, T. P. Catastrophes, phase shifts and large-scale degradation of a Caribbean coral reef. *Science* **265**, 1547–1551 (1994).
96. Mumby, P. J., Hastings, A. & Edwards, H. J. Thresholds and the resilience of Caribbean coral reefs. *Nature* **450**, 98–101 (2007).
97. Hicks, C. C., Crowder, L. B., Graham, N. A. J., Kittinger, J. N. & Le Cornu, E. Social drivers forewarn of marine regime shifts. *Front. Ecol. Environ.* **14**, 252–260 (2016).
98. Cinner, J. E. et al. Linking social and ecological systems to sustain coral reef fisheries. *Curr. Biol.* **19**, 206–212 (2009).
99. O'Neill, D. W., Fanning, A. L., Lamb, W. F. & Steinberger, J. K. A good life for all within planetary boundaries. *Nat. Sustain.* **1**, 88–95 (2018).
This paper outlines the sustainability challenges of current development trajectories.
100. Stern, D. I., Common, M. S. & Barbier, E. B. Economic growth and environmental degradation: the environmental Kuznets curve and sustainable development. *World Dev.* **24**, 1151–1160 (1996).
101. Alamgir, M. et al. Economic, socio-political and environmental risks of road development in the tropics. *Curr. Biol.* **27**, R1130–R1140 (2017).
102. Deininger, K. & Byerlee, D. *Rising Global Interest in Farmland* (The World Bank, Washington, D.C., 2011).
103. Laurance, W. F., Sayer, J. & Cassman, K. G. Agricultural expansion and its impacts on tropical nature. *Trends Ecol. Evol.* **29**, 107–116 (2014).
104. Pauly, D. On Malthusian overfishing. *Naga, the ICLARM Quarterly* **13**, 3–4 (1990).
105. Rands, M. R. W. et al. Biodiversity conservation: challenges beyond 2010. *Science* **329**, 1298–1303 (2010).
This study shows that effective environmental governance is a necessary condition for improved sustainability outcomes.
106. Blasiak, R. & Wabnitz, C. C. C. Aligning fisheries aid with international development targets and goals. *Mar. Policy* **88**, 86–92 (2018).
107. Mora, C. et al. Management effectiveness of the world's marine fisheries. *PLoS Biol.* **7**, e1000131 (2009).
108. Mammides, C. et al. Increasing geographic diversity in the international conservation literature: a stalled process? *Biol. Conserv.* **198**, 78–83 (2016).
This paper reveals that authors from low-income countries are less than half as likely to be published as those from high-income countries.
109. Lovejoy, T. E. & Nobre, C. Amazon tipping point. *Sci. Adv.* **4**, eaat2340 (2018).
110. Nilsson, M. *Important Interactions among the Sustainable Development Goals Under Review at the High-Level Political Forum 2017* (SEI, Stockholm, 2017).
111. Kopnina, H., Washington, H., Gray, J. & Taylor, B. The 'future of conservation' debate: defending ecocentrism and the Nature Needs Half movement. *Biol. Conserv.* **217**, 140–148 (2018).
112. UNEP-WCMC & IUCN. Protected Planet: the World Database on Protected Areas (WDPA) https://www.protectedplanet.net/c/world-database-on-protected-areas (2018).
113. Watson, J. E. M., Dudley, N., Segan, D. B. & Hockings, M. The performance and potential of protected areas. *Nature* **515**, 67–73 (2014).
114. Büscher, B. et al. Half-Earth or Whole Earth? Radical ideas for conservation, and their implications. *Oryx* **51**, 407–410 (2017).
115. Symes, W. S., Rao, M., Mascia, M. B. & Carrasco, L. R. Why do we lose protected areas? Factors influencing protected area downgrading, downsizing and degazettement in the tropics and subtropics. Glob. *Chang. Biol.* **22**, 656–665 (2016).
116. Agrawal, A., Nepstad, D. & Chhatre, A. Reducing emissions from deforestation and forest degradation. *Annu. Rev. Environ. Resour.* **36**, 373–396 (2011).
117. Winder, G. M. & Le Heron, R. Assembling a Blue Economy moment? Geographic engagement with globalizing biological–economic relations in multi-use marine environments. *Dialogues Hum. Geogr.* **7**, 3–26 (2017).
118. Lambin, E. F. et al. The role of supply-chain initiatives in reducing deforestation. *Nat. Clim. Chang.* **8**, 109–116 (2018).
This study highlights the importance of tackling demand for unsustainable products from downstream buyers and investors.
119. Carlson, K. M. et al. Effect of oil palm sustainability certification on deforestation and fire in Indonesia. *Proc. Natl Acad. Sci. USA* **115**, 121–126 (2018).
120. Richards, R. C. et al. Governing a pioneer program on payment for watershed services: stakeholder involvement, legal frameworks and early lessons from the Atlantic forest of Brazil. *Ecosyst. Serv.* **16**, 23–32 (2015).
121. Kosoy, N. & Corbera, E. Payments for ecosystem services as commodity fetishism. *Ecol. Econ.* **69**, 1228–1236 (2010).
122. Corbera, E. & Schroeder, H. REDD+ crossroads post Paris: Politics, lessons and interplays. *Forests* **8**, 508 (2017).
123. Liu, J. et al. Systems integration for global sustainability. *Science* **347**, 1258832–1258832 (2015).
124. Cinner, J. E. et al. Comanagement of coral reef social-ecological systems. *Proc. Natl Acad. Sci. USA* **109**, 5219–5222 (2012).
125. Porter-Bolland, L. et al. Community managed forests and forest protected areas: an assessment of their conservation effectiveness across the tropics. *For. Ecol. Manage.* **268**, 6–17 (2012).
126. Gudynas, E. Buen Vivir: today's tomorrow. *Development* **54**, 441–447 (2011).
127. Seymour, F. & Busch, J. *Why Forests? Why Now? The Science, Economics, and Politics of Tropical Forests and Climate Change* (Center for Global Development, Washington, D.C., 2016).

© 2018 Springer Nature Limited. All rights reserved.

128. Mace, G. M. Whose conservation? *Science* **345**, 1558–1560 (2014).
129. Soulé, M. in *Keeping the Wild: Against the Domestication of Earth* (eds Wuerthner, G. et al.) 66–80 (Island, Washington, D.C., 2014).
130. Holmes, G., Sandbrook, C. & Fisher, J. A. Understanding conservationists' perspectives on the new-conservation debate. *Conserv. Biol.* **31**, 353–363 (2017).
131. Matulis, B. S. & Moyer, J. R. Beyond inclusive conservation: the value of pluralism, the need for agonism, and the case for social instrumentalism. *Conserv. Lett.* **10**, 279–287 (2017).
This paper highlights the importance of pluralism in conservation approaches.
132. Montesino Pouzols, F. et al. Global protected area expansion is compromised by projected land-use and parochialism. *Nature* **516**, 383–386 (2014).
133. Larsen, R. K. et al. Hybrid governance in agricultural commodity chains: insights from implementation of 'No Deforestation, No Peat, No Exploitation' (NDPE) policies in the oil palm industry. *J. Clean. Prod.* **183**, 544–554 (2018).
134. Winemiller, K. O. et al. Balancing hydropower and biodiversity in the Amazon, Congo, and Mekong. *Science* **351**, 128–129 (2016).
135. van Oppen, M. J. H., Oliver, J. K., Putnam, H. M. & Gates, R. D. Building coral reef resilience through assisted evolution. *Proc. Natl Acad. Sci. USA* **112**, 2307–2313 (2015).
136. Svenning, J.-C. et al. Science for a wilder Anthropocene: synthesis and future directions for trophic rewilding research. *Proc. Natl Acad. Sci. USA* **113**, 898–906 (2016).
137. Hortal, J. et al. Seven shortfalls that beset large-scale knowledge of biodiversity. *Annu. Rev. Ecol. Evol. Syst.* **46**, 523–549 (2015).
138. DRYFLOR et al. Plant diversity patterns in Neotropical dry forests and their conservation implications. *Science* **353**, 1383–1387 (2016).
139. Leal, C. G. et al. Is environmental legislation conserving tropical stream faunas? A large-scale assessment of local, riparian and catchment-scale influences on Amazonian fish. *J. Appl. Ecol.* **55**, 1312–1326 (2018).
140. Pitman, N. C. A., Azáldegui, M. d. C. L., Salas, K., Vigo, G. T. & Lutz, D. A. Written accounts of an Amazonian landscape over the last 450 years. *Conserv. Biol.* **21**, 253–262 (2007).
141. Feeley, K. Are we filling the data void? An assessment of the amount and extent of plant collection records and census data available for tropical South America. *PLoS ONE* **10**, e0125629 (2015).
142. Sutherland, W. J., Gardner, T. A., Haider, L. J. & Dicks, L. V. How can local and traditional knowledge be effectively incorporated into international assessments? *Oryx* **48**, 1–2 (2014).
143. Ghazoul, J. & Chazdon, R. Degradation and recovery in changing forest landscapes: a multiscale conceptual framework. *Annu. Rev. Environ. Resour.* **42**, 161–188 (2017).
144. Liu, J. et al. Framing sustainability in a telecoupled world. *Ecol. Soc.* **18**, art26 (2013).
145. Brashares, J. S. Bushmeat hunting, wildlife declines, and fish supply in West Africa. *Science* **306**, 1180–1183 (2004).
146. Daskin, J. H. & Pringle, R. M. Warfare and wildlife declines in Africa's protected areas. *Nature* **553**, 328–332 (2018).
147. Richards, P. D., Myers, R. J., Swinton, S. M. & Walker, R. T. Exchange rates, soybean supply response, and deforestation in South America. *Glob. Environ. Change* **22**, 454–462 (2012).
148. Hicks, C. C. et al. Engage key social concepts for sustainability. *Science* **352**, 38–40 (2016).
149. Fischer, J. et al. Advancing sustainability through mainstreaming a social–ecological systems perspective. *Curr. Opin. Environ. Sustain.* **14**, 144–149 (2015).
This paper examines how the concept of social–ecological systems can improve sustainability.
150. Cvitanovic, C. et al. Improving knowledge exchange among scientists and decision-makers to facilitate the adaptive governance of marine resources: a review of knowledge and research needs. *Ocean Coast. Manage.* **112**, 25–35 (2015).
151. Cinner, J. E. et al. Bright spots among the world's coral reefs. *Nature* **535**, 416–419 (2016).
This analysis shows how the 'bright spots' approach can identify better-than-expected environmental situations.
152. Bennett, E. M. et al. Bright spots: seeds of a good Anthropocene. *Front. Ecol. Environ.* **14**, 441–448 (2016).
153. Pringle, R. M. Upgrading protected areas to conserve wild biodiversity. *Nature* **546**, 91–99 (2017).
154. Balvanera, P. et al. Key features for more successful place-based sustainability research on social-ecological systems: a Programme on Ecosystem Change and Society (PECS) perspective. *Ecol. Soc.* **22**, 14 (2017).
155. Clarke, D. A., York, P. H., Rasheed, M. A. & Northfield, T. D. Does biodiversity–ecosystem function literature neglect tropical ecosystems? *Trends Ecol. Evol.* **32**, 320–323 (2017).
156. Gardner, T. A. et al. A social and ecological assessment of tropical land uses at multiple scales: the Sustainable Amazon Network. *Phil. Trans. R. Soc. B* **368**, 20120166 (2013).
157. Rose, R. A. et al. Ten ways remote sensing can contribute to conservation. *Conserv. Biol.* **29**, 350–359 (2015).
158. Thomsen, P. F. & Willerslev, E. Environmental DNA – an emerging tool in conservation for monitoring past and present biodiversity. *Biol. Conserv.* **183**, 4–18 (2015).
159. Gardner, T. A. et al. Transparency and sustainability in global commodity supply chains. *World Dev.* (2018)
160. Basset, Y. et al. Conservation and biological monitoring of tropical forests: the role of parataxonomists. *J. Appl. Ecol.* **41**, 163–174 (2004).
161. Barlow, J. et al. Using learning networks to understand complex systems: a case study of biological, geophysical and social research in the Amazon. *Biol. Rev. Camb. Philos. Soc.* **86**, 457–474 (2011).
162. Barlow, J. et al. On the extinction of the single-authored paper: the causes and consequences of increasingly collaborative applied ecological research. *J. Appl. Ecol.* **55**, 1–4 (2018).
163. Dinerstein, E. et al. An ecoregion-based approach to protecting half the terrestrial realm. *Bioscience* **67**, 534–545 (2017).
164. Kleypas, J. A., McManus, J. W. & Meñez, L. A. B. Environmental limits to coral reef development: where do we draw the line? *Am. Zool.* **39**, 146–159 (1999).
165. Birdlife International & Handbook of the Birds of the World. Bird species distribution maps of the world Version 7.0. http://datazone.birdlife.org/species/requestdis (2017).
166. Tedesco, P. A. et al. A global database on freshwater fish species occurrence in drainage basins. *Sci. Data* **4**, 170141 (2017).
167. Cochrane, M. A. Fire science for rainforests. *Nature* **421**, 913–919 (2003).
168. Flores, B. M., Fagoaga, R., Nelson, B. W. & Holmgren, M. Repeated fires trap Amazonian blackwater floodplains in an open vegetation state. *J. Appl. Ecol.* **53**, 1597–1603 (2016).
169. Jolly, W. M. et al. Climate-induced variations in global wildfire danger from 1979 to 2013. *Nat. Commun.* **6**, 7537 (2015).
170. Durigan, G. & Ratter, J. A. The need for a consistent fire policy for Cerrado conservation. *J. Appl. Ecol.* **53**, 11–15 (2016).

Acknowledgements We thank the following for providing data: Birdlife International, International Union for Conservation of Nature, Ocean Biogeographic Information System, C. Veron for zooxanthellate corals, the Large-Scale Biosphere-Atmosphere Program (LBA) and National Environment Research Council grant NE/P004512/1 for forest birds, John Fell Fund for savannah birds, Pew Marine Fellows Program of The Pew Charitable Trusts, and Brazilian Council for Scientific and Technological Development (CNPq) (PELD 441659/2016-0). Individual funding was provided by National Environment Research Council NE/K016431/1; NE/P004512/1; NE/L000016/1; European Research Council 759457; H2020-MSCA-RISE (691053-ODYSSEA); NASA's Interdisciplinary Research in Earth Sciences program (NNX14AD29G); OIST and JSPS KAKENHI (JP17K15180); CNPq scientific productivity grant (307788/2017-2); CNPq Programa de Capacitação Institucional (300231/2016-4); Royal Society (UF140691) and Swedish Research Council Formas grant (2016-00351).

Reviewer information *Nature* thanks N. Knowlton, C. Mora and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.B. developed the review with input from N.A.J.G., T.A.G., C.C.H., A.C.L. and J.F. F.F. and G.D.L. analysed the data, supported by J.B., T.A.G., C.C.H., E.B., L.C., E.P.E., B.G., C.G.L., V.I., A.C.L., C.L.P., S.K.W., P.J.Y. and N.A.J.G. J.B., T.A.G., C.C.H., N.A.J.G., L.C., A.C.L., C.L.P., F.F. and G.D.L. wrote the manuscript with input from all authors.

Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0301-1.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.B.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

REVIEW

https://doi.org/10.1038/s41586-018-0300-2

The tropical forest carbon cycle and climate change

Edward T. A. Mitchard[1]*

Tropical forests make an approximately neutral contribution to the global carbon cycle, with intact and recovering forests taking in as much carbon as is released through deforestation and degradation. In the near future, tropical forests are likely to become a carbon source, owing to continued forest loss and the effect of climate change on the ability of the remaining forests to capture excess atmospheric carbon dioxide. This will make it harder to limit global warming to below 2 °C. Encouragingly, recent international agreements commit to halting deforestation and degradation, but a lack of fundamental data for use in monitoring and model design makes policy action difficult.

Tropical forests have had a critical role in the changing atmospheric carbon concentrations of the industrial age. They have acted both as a very important source of emissions, as a result of logging and burning, and also as a carbon sink, as the remaining forests have taken in much of the extra carbon added to the atmosphere. As an example, from 1960 to 2015, anthropogenic emissions totalled 408 Pg carbon (Pg C), of which 80% arose from burning fossil fuels and the production of cement, and 20% from changes in (largely tropical) land use[1]. However, the atmospheric CO_2 stock grew by 'only' 180 Pg C over that period[2,3], meaning that 55% of these emissions were taken in by the Earth system, reducing the magnitude of the induced climate change. This sink has three main components—the oceans, Northern Hemisphere forests and tropical forests[1]—and although there are uncertainties around the relative contributions of each, between a quarter and a third is normally attributed to the enhanced growth of trees in tropical forests[4–6]. Understanding the size and the causes of this sink is crucial for predicting its evolution over the coming century: the tropical land sink is known to be very variable year to year[1,7–10], and reverses to become a source in hotter years[11]. This suggests there is a real risk that, over the coming decades under climate change, it will become a major source of carbon every year.

The tropical land sink is the least certain major component of the global carbon budget[1]. There are various possible ways of estimating its size (Box 1), but none estimate the sink directly, and all have high uncertainty due to either sparse sampling[6,12–16] or coarse resolution[7,8,11]. As a result, the main way in which the land sink has been estimated is as the residual of the sum of all other components of the global carbon cycle[1]; however, with this method it is not possible to estimate the relative contributions of the Northern Hemisphere forests and the tropical forests to the sink. In addition, some other components of the global carbon cycle are also very uncertain and variable, such as the land-use change flux[1,17], making the accurate estimation of trends in the sink very difficult. This uncertainty greatly limits the development and testing of theories and models, and thus means that there is a wide divergence of predictions as to how the sink will change under different climate-change scenarios and policy interventions.

Considering all sources of evidence, it appears probable that as the intact forest sink declines in size, tropical forests are in the process of switching from being approximately neutral to being a net source of carbon[1,4,8,18,19]. This decline is caused by the combination of a decrease in the area of intact forest[20,21] as well as an increase in temperatures and drought, which reduces the ability of trees to respond to higher CO_2 concentrations by growing faster[13,14]. With both forest loss and climate change likely to accelerate over the 21st century, tropical forests are likely to release ever more carbon, which will make limiting global warming to less than 2 °C above pre-industrial levels very difficult[22,23].

The carbon balance of tropical forests

Living tropical trees store 200–300 Pg C (refs [6,24–26]), about a third as much as is held in the atmosphere[1]. This stock is very dynamic: tropical trees perform about 60% of the world's photosynthesis, capturing around 72 Pg C from the atmosphere every year[27], but also release a similar amount back to the atmosphere through respiration of both the plants themselves and the animals, microorganisms and fungi of the ecosystem[18,28]. Given these large fluxes, a small proportional change in either the uptake or the release of CO_2 can result in a large net source or sink. There are several lines of evidence that, over at least the past 50 years, these two processes have been out of balance, with tropical vegetation increasing in biomass by more than 2 Pg C yr^{-1}, equivalent to about 1% per year[6,19,29,30]. However, it is clear that the interannual variability of this sink is very high, driven by fluctuations in temperature and rainfall[17,31,32].

The tropics are also the main nexus of global land-use change, with deforestation and forest degradation (where some trees are removed but the area retains sufficient trees to be classed as a forest) releasing somewhere between 0.5 and 3.5 Pg C yr^{-1} (refs [8,21,33–38]). The wide range of estimates can be partly attributed to differences in time period, but is mostly caused by differing definitions and included processes, different methods (Box 1) and wide uncertainty bounds.

The comparison of different methods shows a consensus that the overall carbon balance of the tropics has been approximately neutral since at least the 1990s, with sinks in intact and regrowing forests equal in magnitude to sources from deforestation and forest degradation[5,6,8] (Fig. 1). However, it is also clear that in abnormally hot years, such as during strong El Niño events, the tropics become a major net source[1,11] (Fig. 1d).

The following section examines the current magnitude of the major carbon sources and sinks. Next, the evidence for trends in the sources and sinks over recent decades is considered, along with their probable future pathways, and whether international policy can change these trends.

Carbon sources

Deforestation is easy to map using optical satellite data. Free Landsat satellite data enable most countries to produce their own maps[21];

[1]School of GeoSciences, University of Edinburgh, Edinburgh, UK. e-mail: edward.mitchard@ed.ac.uk

© 2018 Springer Nature Limited. All rights reserved.

Box 1

Methods used to assess the tropical forest carbon balance

Forest inventory plots. Marked areas of forest where tree diameters are measured and species recorded, enabling estimation of tree mass[89]. Revisiting networks of such plots approximately every five years gives precise estimates of how forest carbon stocks are changing[13,14,57], although uncertainties are increased because plots are rare and unevenly distributed, with some forest types undersampled. There are also plots that are intensively monitored to give insight into the detail of carbon allocation and use efficiency[90], and rare experimental manipulations that test the response of trees to conditions that do not naturally exist[75,91].

Atmospheric inversions. There is a sparse network of towers and marine measurement sites across the tropics that permanently collect greenhouse gas concentration and micrometeorological data. These are supplemented by ship and aircraft data, and combined with atmospheric transport models to estimate the net flows of CO_2 into or out of the atmosphere at a broad, regional scale[5,15,16].

Satellites. Can be used to estimate:

(1) Forest area. Landsat satellites have been used to produce consistent estimates of forest cover change since the early 1970s. Many countries produce their own maps, and global 30-m resolution forest change data are available from 2000 onwards[20]. However, loss data are much more reliable than gain data, and relating the area-based data to changes in carbon stock is difficult.

(2) Carbon stocks. A unique LiDAR satellite operating in the mid-2000s collected distributed estimates of tree height in 70-m footprints, which were combined with field plots and other satellite data to make medium-resolution (500 m–1 km pixels) carbon stock maps[24,25], albeit with large uncertainties[40]. These maps enable estimates of emissions when combined with forest area change data[36], or when produced annually[8]. It is also possible to estimate carbon stock changes using passive microwave remote sensing[37], however the resolution (1–2 times coarser) makes it impossible to separate gain and loss fluxes.

(3) GHG concentration. Satellites can measure the greenhouse gas concentrations in narrow columns of the atmosphere with a precision of approximately 1 p.p.m. CO_2. These measurements have been used to observe directly the carbon entering and leaving tropical forests, giving information about the size of the tropical forest sink and its reaction to droughts at a continental scale[7,11]. However, cloud cover means that the observations are sparse in time and space, and the coarse resolution once again means forest loss and gain fluxes cannot be distinguished.

Modelling. Given the difficulty in directly observing forest responses to increasing CO_2 concentrations and climate changes, dynamic vegetation models are used directly to predict these responses[9,18]. The latest generation of models, Earth system models (ESMs), include many more processes and feedbacks than do traditional atmosphere–ocean general circulation models (AOGCMs), which increases their predictive power[71]. Models provide information on processes or time periods for which we have no other data, and enable us to use our current knowledge about the Earth system to predict the future under specific scenarios of climate and land-use change[23].

large-scale independent maps are also available and are broadly consistent[20,39]. Deforestation affects very large areas: about 100 Mha was deforested in the tropics from 2000–2012; of this, about 50% was in Latin America, 30% in southeast Asia and 20% in Africa (ref. [20], using a forest definition of greater than 25% canopy cover; similar values are found in other studies[21,39]). The main drivers of this deforestation differ by location: mining and large-scale commercial agriculture or pasture dominate in Latin America; palm oil, pulp and paper plantations in southeast Asia; whereas in Africa, deforestation has been largely driven by smallholder agriculture, with major contributions only more recently from mining, commercial agriculture and plantations[33].

Estimating the amount of carbon released from deforestation is more difficult than assessing its spatial extent. Estimates are often produced by simply multiplying the area deforested by a single carbon density per unit area value. The result is therefore very sensitive to that single value, which is usually the mean carbon density from a number of local forest inventory plots; however, to be accurate they must be numerous and representative of the type of forest removed. At a pantropical scale, recent studies have improved on this by overlaying the deforestation data on continuous maps of carbon density[8,33,36,38]. However, errors can arise from the carbon data having a coarser resolution than the deforestation data, and the carbon maps having potentially large regional biases[26,40]. Overcoming these issues, there has been some consensus in recent years that the flux from gross tropical deforestation in the 2000s was 0.6–0.8 Pg C yr^{-1} (refs [33,36]).

By contrast, it is much harder to estimate the area affected and the carbon losses caused by forest degradation[41,42]. This is partly because degradation is caused by a wide variety of processes with different effects, including commercial logging, fuelwood extraction, sub-canopy cultivation, grazing, fire, and edge effects caused by nearby deforestation[41,43]. Further, it is because the only remote sensing methods that are sensitive to degradation have coarse resolution, with each pixel containing between twenty hectares and thousands of hectares[5,7,8,11,15,16,37,44], which far exceeds the size of most degradation events (less than 1 ha)[42,45]. This means that estimates inevitably mix the fluxes from deforestation, forest degradation, regrowth of previously disturbed forest, as well as changes in intact forest, into a single combined change per pixel.

Some studies have used inventory plots to estimate fluxes from degradation[41,46]. However, these give numbers on a per-hectare basis that are hard to scale, as we do not have maps of degradation. High-resolution remote sensing from light detection and ranging (LiDAR)[47] or synthetic-aperture radar[45], combined with local field biomass plots, can directly map the changes in carbon stock from deforestation, degradation and regrowth at a suitable resolution, but so far such studies are rare and have been used only for small areas, so cannot be of much help for pantropical estimates. They can, however, show the broad ratio between carbon losses from deforestation and degradation; although this varies widely in space and time, there is a suggestion that, at a large scale, degradation is responsible for perhaps twice the carbon release of deforestation, with great regional variation[8,45]. Furthermore, there is agreement that degradation is more important as a proportion of total emissions in Africa than in South America or southeast Asia[30,41,42,48].

Tropical peat forests are independently a major potential source of carbon. Peat is carbon-rich, partially decayed organic matter, associated with waterlogged and acidic conditions, which exists in layers up to 20-m thick under tropical swamp forests. Recent large discoveries under the forests of the Congo[49] and Amazon[50] basins have increased the known area of tropical peat by 50% to 577,000 km^2 (combining figures from refs [49,51]). These peat forests have very high carbon densities, meaning that they have the potential to make an outsized contribution to the global carbon cycle: about 5% of tropical forests overlay peat, but they store 70–130 Pg C (ref. [49]), which is considerable compared to the 200–300 Pg C stored in all tropical trees[6,24–26]. The majority of tropical peat is in southeast Asia, but it has been extensively cleared and drained in recent decades (over half the area present in 1990 had been deforested or degraded by 2008[52,53]), therefore contributing considerably to land-use change emissions by releasing 0.3–0.54 Pg C yr^{-1} (refs [4,17]). This large flux is included in the land-use change numbers in Fig. 1, but is excluded from normal values of the deforestation flux (for example, refs [8,33,36,38]), as such methods exclude belowground carbon.

© 2018 Springer Nature Limited. All rights reserved.

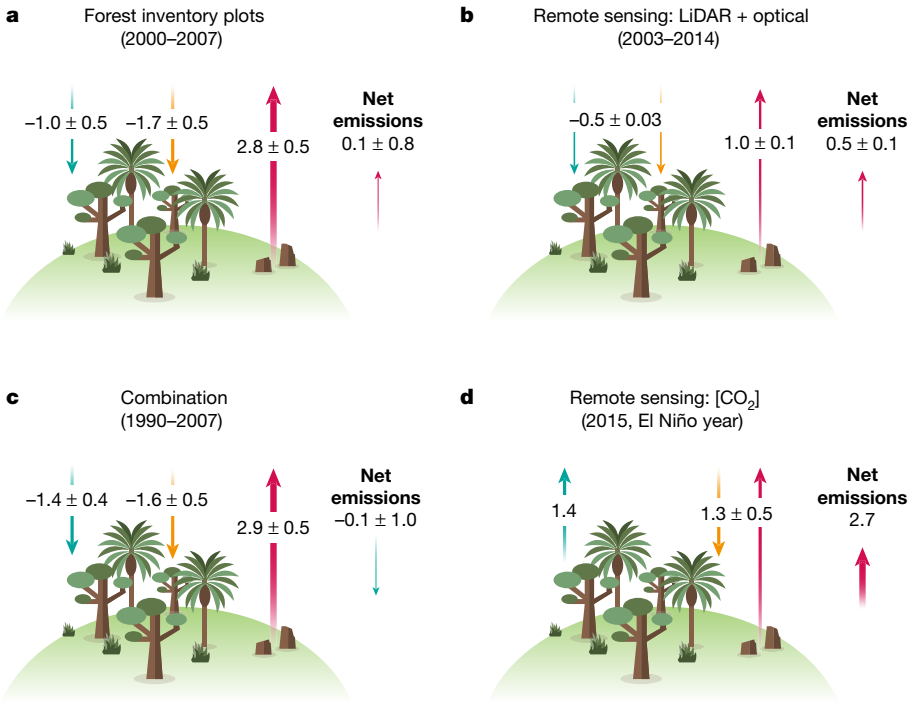


Fig. 1 | Tropical forest carbon fluxes assessed using different methods. Annual fluxes (in Pg C yr^{-1}) into and out of tropical forests for different overlapping time periods (**a**–**c**) and for a recent El Niño year (**d**). The net intact forest flux is shown in turquoise, the net flux in regrowing forest is in orange, and the deforestation and forest degradation flux (including fire) is shown in pink. Panels **a**–**c** show that there is broad agreement that the tropics have made an approximately neutral contribution to atmospheric carbon stocks in the recent past, but **d** shows that intact forest can become a carbon source in hot and dry years, leading to considerable net emissions from the tropics. Data in **a** are from networks of forest inventory plots[6], combined with forest area data from country surveys or ref. [48]. In **b**, values are obtained from annual 463-m resolution optical satellite data[8], calibrated using LiDAR data and field plots from the mid-2000s. Intact and regrowth fluxes are not separated in this method. The figures in the study have been grossed up from biomass to total carbon stock change (that is, including dead wood, litter, soil) using the data in table 2 of ref. [6] (adding 16%). Data in **c** are derived from looking for overlap between atmospheric inversion, modelling and field-plot estimates[5]. In **d**, data are obtained from satellites sensitive to atmospheric CO_2 concentrations for the 2015 El Niño year[11], which contrast sharply with the other estimates shown. Change in land use could not be divided into separate regrowth and loss fluxes in this method.

Furthermore, intact, degraded and drained peatlands in southeast Asia have been subject to fires in El Niño years that have released much larger quantities of carbon: up to 2.5 Pg C in a single year, sufficient to cause noticeable anomalies in the atmospheric CO_2 growth rates[11,54]. By contrast, the peatlands of the Congo and Amazon basins were until recently largely undisturbed, and so are not currently thought to contribute considerably to the land-use change flux[55].

However, even undisturbed peatlands are probably losing carbon owing to climate change[4,51,53]. This is hard to monitor because satellites can see only the trees, but the vast majority of the carbon in peat-forest ecosystems is instead stored belowground as peat. Gruelling fieldwork to ascertain peat depths and extract cores for chemical analysis is necessary to ascertain carbon stocks, but these point estimates are hard to scale to large areas owing to great spatial variability[49,53]. The tracking of losses is further complicated because of the range of mechanisms through which peats can lose carbon: respiration in peats releases CH_4 as well as CO_2; burning releases CO and C in addition to CO_2; and dissolved and particulate organic carbon is washed away in rivers. There is some data on non-CO_2 emissions: both satellite and modelling datasets suggest that all tropical peatlands are substantial methane sources[56], and field data suggests that fluvial organic carbon transport is considerable for both intact and disturbed peats in southeast Asia, and has increased by more than 30% from 1990 to 2008[52]. Although more baseline data are needed, it seems likely that warming and droughts caused by climate change are resulting in peat forests becoming net sources of carbon[4,19,51,53].

Carbon sinks

The remeasurements of millions of trees in networks of forest inventory plots across the undisturbed forests of Latin America, Africa and southeast Asia suggest that these forests have all been gaining carbon at a similar rate of around 0.5 Mg C ha^{-1} yr^{-1} since at least the late 1970s, adding up to a total sink of a little over 1 Pg C yr^{-1} (refs [6,14,57,58]). Although it has been suggested that artefacts in plot remeasurements could lead to erroneous findings of increasing carbon storage with time[59], there is also considerable evidence of a sink of around this magnitude from independent methods, such as atmospheric inversion studies[5,15,16], satellite data[8,37] and models[18], so there is little doubt that it exists.

Regrowing and disturbed forests are also clearly taking in carbon from the atmosphere, but as with forest degradation, there is little reliable data on the magnitude of this sink. Studies that track individual field plots show great variation: after total clearance there was no increase in biomass at all after 20 years at a site in Uganda[60], but an increase of over 10 Mg C ha^{-1} yr^{-1} throughout the first 10 years in moist sites in Latin America[12]. A meta-analysis of 1,468 plots in 45 sites found average recovery rates of 3.05 Mg C ha^{-1} yr^{-1} for the first 20 years, and that sites regained 90% of old-growth biomass values after a median of 66 years[12] (though biodiversity does not recover in these timescales[12,42,60]). As we do not have good maps of past disturbance it is hard to convert plot values into tropical estimates, but these data are consistent with inversion studies and satellite observations of a current flux with a similar magnitude to the intact forest sink (that is, 1 Pg C yr^{-1}), with large uncertainty[4–6].

Trends in the sources and sinks

Despite the uncertainties related to individual components, we do have a reasonable understanding of the carbon balance of the tropics in the recent past, with various methods agreeing that the tropics make an approximately neutral contribution to the global carbon budget[4,5,8,37] (Fig. 1). However, we are much less certain about how the system is changing.

© 2018 Springer Nature Limited. All rights reserved.

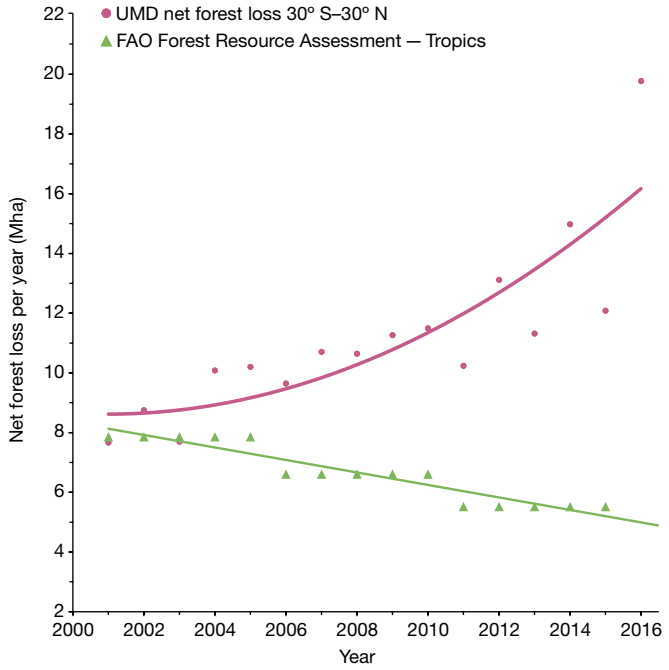


Fig. 2 | Contradiction between the main datasets of forest area change. Satellite datasets and nationally reported statistics are in agreement about the rate of net tropical forest loss in the early 2000s, but diverge increasingly over time. The pink points and trendline (quadratic ordinary least squares) show net annual forest loss from a systematic global satellite analysis from the University of Maryland (UMD)[20], Version 1.4. Forest gain is not assigned to a particular year in this dataset, so is here distributed equally across the time period to give net figures. A forest definition of 10% canopy cover in 2000 was used. The green points and trendline (linear ordinary least squares) show net annual change in forest area across tropical countries from the FAO Global Forest Resources Assessment 2015[48], as summarized in ref. [21]. Forest area is reported at five-year intervals, the change has been calculated between each interval and then divided by five to give annual data. Various forest definitions are used by countries when producing these figures, with canopy cover ranging from 10–30%. This means that the total area of forest considered for 2000 is higher in the UMD dataset, and would be expected to lead consistently to slightly higher deforestation for the UMD dataset compared to the FAO dataset. However, this difference in forest definition cannot explain the differences in trend, as only around 5% of losses in the UMD dataset are in forest with canopy cover of between 10 and 30%.

Although there is general agreement that total forest area is shrinking across the tropics, there is considerable controversy as to whether the rate of loss is increasing or decreasing. Official figures from the Food and Agriculture Organization (FAO; collated from national statistics) show a decline in annual net forest loss rates since 2000[21,48], whereas satellite-based data show an increase in the loss rate[20] (Fig. 2). Some of this difference can be explained by differing definitions of forest and the precise area compared, but the difference in trend is too large to be explained by these alone. It has long been known that FAO statistics are not ideal for analysing trends[61]: although some tropical countries probably produce very good data, their monitoring capacity is variable[62]. As an example, 14 African countries have reported exactly the same annual change in forest area every year from 1990–2015[48], even though other datasets see considerable changes in their rates of loss over time[34].

On balance, the evidence from the remote-sensing data sources appears to be more reliable than the FAO data. For example, the satellite-derived product from Hansen et al.[20] is produced consistently across the tropics, detects correct country-level trends where we have good alternative sources of data (for example, correctly seeing the rapid reduction in deforestation in Brazil[63] and the recent rapid acceleration of loss in the Democratic Republic of the Congo[64]), and matches well to detailed high-resolution data in a study that compares areas with different patterns of forest loss[65]. It is therefore likely that the rate of deforestation in the tropics is increasing.

Over the coming decades, as the global demand for agricultural, timber and mineral commodities, as well as local population density, continue to grow, it seems likely that the rate of forest loss will continue to increase[33,63,64]. Current areas that are largely undisturbed owing to inaccessibility, such as the peat forests of the western Amazon and central Congo basins, are likely to become accessible and suffer deforestation[55]. Eventually the rate of forest loss will stabilize and start to fall, partly because the area of remaining unprotected forest will have greatly decreased, but also because once countries reach a sufficient level of economic development and forest loss, policy and civil society drivers result in the remaining forest area stabilizing or even increasing[66]. However, this point may come only once most of the forest has been lost, and even when countries reach this point (such as Vietnam, China or much of the developed world), as their economies demand increasing levels of commodities, they themselves will export deforestation to less economically developed countries[67]. The case of Brazil makes an interesting case study here: it greatly reduced its rate of forest loss from 2005–2014[20], owing to reductions in global commodity prices and policy interventions[63], but the rate has since increased again and could climb faster as the global demand for agricultural and mineral commodities increases, and laws promote development and not forest protection[63,68].

Forest degradation is difficult to map and monitor: as discussed previously, there is little hard evidence of its overall current magnitude, let alone trends, although we suspect it involves a much larger area than deforestation each year[8,44–46]. Normally degradation appears to be closely associated with deforestation[44], and it is reasonable to assume in the future that, as the area of accessible forest increases owing to fragmentation caused by deforestation and road building, the area of forest degraded each year will also increase. About 20% of all tropical forest is now within 100 m of an edge, with 84% of these edges being anthropogenic, and this proportion will continue to increase as more anthropogenic edges are created each year than are closed up[43]. The effect of commercial logging, a major direct cause of degradation but also a driver of increased fragmentation and access roads, also seems likely to increase as ever more logging concessions are granted[41,42]. Increases in fire due to climate change may result in increased degradation as well as increased fragmentation over time[42]. Overall it is hard to believe that the area of forest degraded each year is not increasing, nor that it will stop increasing in the near future.

As the area of degraded forest increases, so does the area of forest with the potential for regrowth; as such, the proportion of the forest sink that comes from previously disturbed forest is likely to increase with time[4]. Disturbed forest normally takes up carbon much faster per hectare than does undisturbed forest, though with high variability[12,69]. Under climate change, the rate of growth that could be achieved by disturbed forest could increase further, owing to CO_2 fertilization[18,70]. However, other factors (such as increased temperature and changing precipitation) could negate this effect: a specific modelling test as to whether land-use change increased the land sink under an extreme CO_2 scenario found that only one out of four models predicted an increased sink, with the others showing no increased sink[70]. Fundamentally, the size of the sink from regrowing forest is very hard to model using current knowledge, with a high level of divergence between models[71]. There is separate evidence that normally, once deforested, land is ultimately permanently converted to agriculture, pasture or settlements, and most degraded land is itself ultimately deforested[72]. This pattern is unlikely to change as the global economy and population continue to grow, so ultimately carbon captured in the regrowing sink may not remain captured for long.

It is also difficult to predict the effect that climate change will have on the intact forest sink, because we know that climate change will bring opposing pressures[6] (Fig. 3). Theory and modelling studies generally agree that the most likely cause of the sink is CO_2 fertilization: as atmospheric CO_2 concentrations have increased from around 280 p.p.m. in 1850 to 410 p.p.m. in mid-2018, fixing carbon through photosynthesis

© 2018 Springer Nature Limited. All rights reserved.

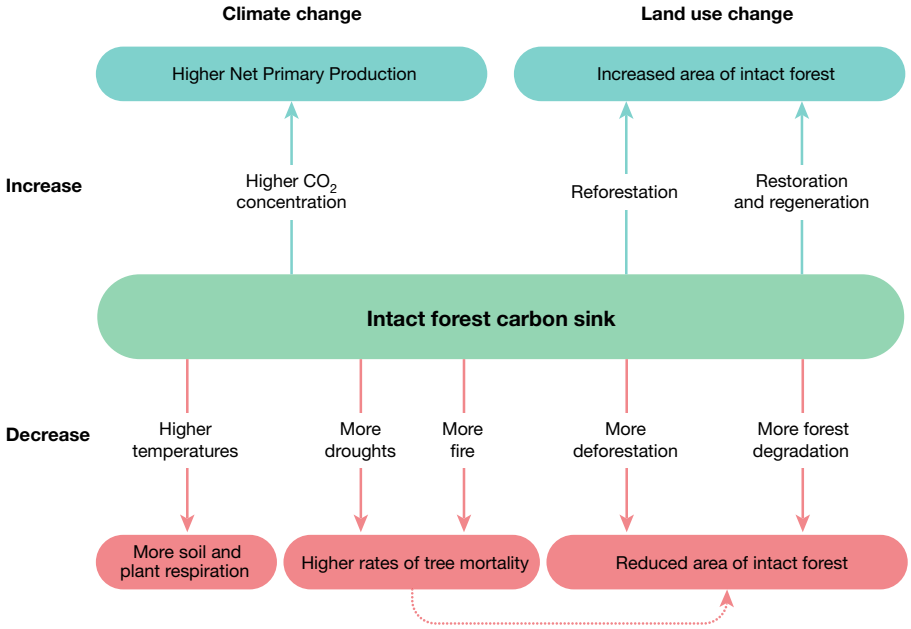


Fig. 3 | The effects of climate and land-use change on the intact forest carbon sink. The potential contrasting effects of climate change and different land-use change trajectories on the size of the intact forest carbon sink. Arrows pointing up show how climate change and policy could increase the magnitude of the sink, whereas arrows pointing down show how it will be reduced. All processes will occur to some extent, so predicting how the sink size will change is very difficult.

has become easier, owing to increased CO_2 concentrations in leaves for a given level of stomatal opening (itself limited by water availability)[18]. This effect should continue as CO_2 levels increase[9,18]; however, climate change will also raise temperatures, which will increase soil and plant respiration rates, and droughts and fires will also likely increase in both frequency and severity, directly killing trees (Fig. 3). In addition, deforestation and degradation will continue to reduce the area of intact forest that can act as a sink. Many studies therefore suggest that climate change could lead to a reduction in the strength of the carbon sink, and ultimately to its reversal into a carbon source[11,73–75]. There is evidence from networks of field plots that this is already happening, with the magnitude of the sink decreasing over time[13,14].

However, models do not generally predict a reduction in the magnitude of the land sink, with many predicting that CO_2 fertilization will offset the negative influence of climate change on ecosystem respiration and tree mortality[9,18,70,76]. For example, six coupled climate models that were run under the same CO_2 growth scenario predicted changes in tropical land carbon storage between 1960 and 2099 that ranged from −11 Pg C to +319 Pg C, with a mean of +172 Pg C (ref. [9]). The differences between these models are mostly attributed to differences in the sensitivity of tropical vegetation to temperature, and the extent to which the temperature increases owing to non-CO_2 forcings (for example, a reduction in aerosol concentrations, or increases in other greenhouse gases), which do not have an associated positive CO_2-fertilization effect[9]. Owing to the variability in model predictions of the current size of the intact forest carbon sink[9,71,77,78], and the lack of critical factors in the models, such as the mortality of large trees caused by droughts[75], it is difficult to use model-based predictions for predicting trends in the forest sink. Therefore, the best evidence is from field plots[13,14] and satellites[7,11], which show that the intact forest sink is weakening and that it becomes a source in unusually hot years. This suggests that it will probably reverse to become a carbon source under climate change.

There is so little data on the carbon balance of intact peat forests that it is hard to speculate on how they are changing with confidence. But it is likely that increasing temperature and variable precipitation has increased the rate of carbon loss, especially when combined with draining and other disturbance, and that such losses are likely to accelerate in the future[49,52,53,55]. However, there are high uncertainties, and an increase in basic observations of peat forests are urgently needed.

Modelling the future of tropical carbon

From recent trends, it appears likely that the current major sources of emissions (deforestation and degradation of forests, including peat forests) will at least stay stable or probably increase over the coming decades[33,63,64], whereas the current sinks (from intact and regrowing forest) will probably reduce, and could reverse and become sources[11,73–75]. Therefore, it is very likely that tropical forests will become a net source of CO_2 to the atmosphere in the near future, if they have not already[4]. However, estimating the size of this tropical source over time is very difficult, even when considering a specific scenario of land use and climate change, owing to complex feedbacks and interactions between different elements of the carbon cycle, climate change, people, policy and the global economy.

The climate modelling community has produced ever more complex models that include the complex feedbacks between land-use change, climate change and intact forest[71]. A noticeable difference between the 4th and 5th Assessment Report of the International Panel on Climate Change (IPCC) is that the latter uses Earth system models (ESMs) for much of its predictions, rather than atmosphere–ocean general circulation models (AOGCMs)[18]. ESMs include all the processes of AOGCMs, but add representations of biogeochemical cycles—including the full carbon cycle—and couple these cycles with other components that allow for feedbacks. For example, deforestation in an AOGCM simulation will increase the atmospheric CO_2 concentration and change the physical properties of the ground surface, but only in an ESM will the smoke and dust released from deforestation, and their subsequent effect on atmospheric chemistry and the rate of photosynthesis of the remaining trees, be modelled[71].

To standardize the inputs for modelling climate change under different scenarios to 2100, the IPCC developed four representative concentration pathways (RCPs) for its 5th Assessment Report[19]. These are trajectories of atmospheric greenhouse gas concentration and consequent radiative forcing (amount of solar energy absorbed by the Earth), and are named after the prescribed radiative forcing in the year 2100 relative to preindustrial levels in W m^{-2}: RCP2.6,

© 2018 Springer Nature Limited. All rights reserved.

RCP4.5, RCP6 and RCP8.5. They are based on underlying assumptions about social and technological development, and the extent to which climate mitigation activities take place: RCP8.5 assumes that annual fossil-fuel emissions increase rapidly until around 2070 before eventually stabilizing, whereas RCP2.6 assumes that emissions peak by 2020 and then decrease rapidly[79]. Ideally it would be possible to provide confident predictions of the size of the forest sinks under these different scenarios, but unfortunately there is still high variability in the predictions of the tropical carbon cycle using ESMs, and these predictions cannot agree as to whether the tropical land surface will gain or lose carbon overall under the different scenarios[9,77]. Much of the uncertainty (around 80%) in tropical land surface prediction is caused not by uncertainty in the scenarios but by differences in model structure[71]; in particular, differences in the predicted effects of specific climate parameters and CO_2 concentration on net primary productivity and vegetation turnover (including structural shifts, wildfires and mortality)[80]. It is therefore urgent that we improve our knowledge of how the components of the tropical carbon cycle function, to better design and test such models.

Effect of policy on tropical forests

The extreme RCPs (RCP2.6 and RCP8.5) assume similar levels of conversion of tropical forest to agriculture[77], with the differences arising largely from the degree of fossil fuel burning. However, in reality, considerable developments in national and international policy since the last IPCC report in 2013 have made the reduction of tropical deforestation and degradation, and the restoration of previously degraded and deforested tropical land, a key pillar of reducing climate change. This is sensible, because unless tropical deforestation and degradation is reversed, the task of halting the rise in atmospheric CO_2 concentrations would involve decarbonizing the global economy at what would probably be an unfeasible rate[22,23], and it offers the possibility of a different path for the tropical carbon cycle rather than continuing current trends.

The Paris Agreement of 2015, which entered into force in 2016 and is now ratified by 178 of the 197 countries of the United Nations Framework Convention on Climate Change, aims to limit increases in global average temperature to "well below 2 °C above pre-industrial levels". It does not specify how this should be reached, but includes a strong statement in Article 5 that countries "should take action to conserve and enhance … forests"[81]. To assist developing countries with meeting Article 5, it "encourages" all countries to engage in REDD+ (defined as 'reducing emissions from deforestation and forest degradation, and the role of conservation, sustainable management of forests and enhancement of forest carbon stocks in developing countries')[81]. The details of financing and monitoring have yet to be agreed, but there is considerable optimism that REDD+ can succeed in increasing the area of forest, and the proportion of it that is undisturbed, compared to business as usual[33,82,83]. As a result of 'REDD Readiness' funding, most tropical countries have increased their capacity to monitor changes in their own forests[62], and are submitting Forest Reference (Emission) Levels—official baselines against which future emissions can be compared—and plans for reducing emissions below these levels if funding is provided.

Although the Paris Agreement is ambitious in overall terms, its proposals on forests lack concrete detail, stating only that countries should "take action". However, there are other international agreements that involve many or most of the same countries and are more specific. For example, the New York Declaration on Forests, signed by 192 organizations including 40 governments in 2014[84], aims to "at least halve the rate of loss of natural forests globally by 2020 and strive to end natural forest loss by 2030". This was ambitious, but some believed it was achievable[33]. More ambitious still, the UN Sustainable Development Goals (SDGs)[85], agreed in 2015, include as Target 15.2 an aim to "by 2020 … halt deforestation". This was included not only because preventing climate change is one of the key aims, but also because healthy tropical forests are important for the achievement of most of the 17 goals[66]. Few believe that deforestation can really be stopped so fast, but these international agreements will spur at least some countries to enact policies to greatly reduce their deforestation rates[33,63].

The New York Declaration on Forests further aims to restore 150 million hectares of currently deforested or degraded land by 2020, and 350 million hectares by 2030. There are worries related to these ambitious targets: there is a risk that natural grasslands will be afforested, leading to a loss of biodiversity and potentially also soil carbon[86], or that agriculture will be displaced by restored forest, leading to more deforestation elsewhere[87]. Also, many countries have not committed to meet their goal solely through natural forest (for example, through leaving degraded land to regenerate naturally, with the greatest ecological and long-term carbon benefits), but instead will plant monocultures of exotic tree species such as teak and rubber. Nonetheless, this overall enthusiasm for the restoration of forests should be positive for tropical carbon storage (Fig. 3), and if the sites are chosen well and the restoration type carefully considered, it could be highly beneficial for people and the environment[22,87].

Looking further into the future, the Paris Agreement mandates that, by the second half of the 21st century, the remaining anthropogenic greenhouse gas emissions will be balanced by sinks[23]. This will require a large programme of capturing carbon directly from the atmosphere and storing it elsewhere[30]. As tropical trees are by far the most efficient carbon-capture method known, a technology known as bioenergy with carbon capture and storage (BECCS) is proposed, which will generate energy through the burning of tropical plantations and store the produced CO_2 below ground[23]. To meet the negative emissions targets needed to limit global warming to below 2 °C by 2100, models suggest it would need to be implemented on an enormous scale (400–800 Mha; for comparison, India covers 329 Mha)[23,30]. This would clearly make the tropics a major carbon sink, but with negative consequences for biodiversity and ecosystem services.

Overall, these agreements are sufficient to change current trends considerably, and if fully implemented would increase the forest carbon storage of the tropics markedly over the coming century. However, meeting the targets would involve drastic and coordinated action from people, policy makers and companies globally[33,63].

Safeguarding tropical forest carbon

The evidence suggests that unless the world makes a coordinated effort to change from its current course, deforestation, degradation and climate change will combine to make the tropics a net source of carbon to the atmosphere over the coming decades. This is despite increasing CO_2 levels making it easier for intact forests to photosynthesize and absorb carbon[9,18]. However, if we were able to stop deforestation and forest degradation, leave currently degraded forests to recover, and reforest, as targeted in international agreements, then tropical forests would instead probably become an important carbon sink, contributing to the Paris Agreement goal of limiting mean global temperature increases to below 2 °C[22,23]. Retaining and restoring these forests would have further immense benefits to human wellbeing, through maintaining their biodiversity and ecosystem services[66]. However, two interconnected problems limit the achievement of these goals. Our spatial information on how forests are changing is poor, and a lack of field experiments means that ESMs cannot predict well how forests will respond to different scenarios of climate and land-use change.

Although we monitor deforestation well, we do not have good data on changes within forests. We have techniques that can observe the integrated carbon flux over large regions, but we have very little knowledge of the size of the individual processes involved (such as degradation, regrowth, or the impact of droughts and fire). This makes it difficult to design and implement policy: for example, no country has reliable baseline figures on their rate of forest degradation[62], so it is difficult to set targets or create policies to reduce degradation under REDD+, and to receive payments even if such policies are successful. This also limits model development and testing.

© 2018 Springer Nature Limited. All rights reserved.

New satellite missions, such as GEDI and OCO-3 (both planned for launch in 2018) and BIOMASS (2021), will help by producing high resolution, globally consistent maps of forest carbon stock changes for the first time. These will not only assist with targeting and monitoring policies, but will also enable us to discover the magnitude of the forest sink at an unprecedented resolution (less than 1 km^2), and how it is influenced by local conditions and climate fluctuations. However, these satellites require pantropical forest inventory and airborne LiDAR data for calibration and validation. REDD+ will assist directly here: already, considerable funding has been spent on designing and setting up monitoring systems and capacity in developing countries[62]. Unfortunately the data collected is rarely made available to the international scientific community, as publishing such data is against the natural instincts of countries, who wish to protect their sovereignty (there are some exceptions: for example, field and LiDAR data from a recent carbon stock map of the Democratic Republic of the Congo is available at http://panda.maps.arcgis.com). Funders and scientists must persuade countries to be more open, or they will not obtain the full benefits from new satellite missions.

Better maps of forest carbon stocks will make a big difference, removing the current wide spread of figures on the carbon fluxes from tropical forests (Figs. 1, 2), supporting REDD+ and other policy efforts to reduce forest loss, and enabling the testing of ESMs and theories as to how tropical forests respond to climate fluctuations and disturbance events. However, these data will not improve our understanding of how forests will respond to climate and CO_2 conditions that do not currently exist, an understanding that is necessary for improving ESMs. For this we need field experiments, such as those that artificially drought, warm or increase the CO_2 concentration of large tropical forest plots. Such experiments are expensive to run and take many years to produce useful results[75], and therefore at present they are almost non-existent in the tropics[88]. Their development should be supported by governments, as without them there will be no data to develop and test the critical next generation of ESMs[78].

Received: 14 December 2017; Accepted: 11 May 2018;
Published online 25 July 2018.

1. Le Quéré, C. et al. Global carbon budget 2016. *Earth Syst. Sci. Data* **8**, 605–649 (2016).
An annually produced analysis of the best evidence for the size and trends of the components of the global carbon cycle.
2. Tans, P. *NOAA/ESRL* https://www.esrl.noaa.gov/gmd/ccgg/trends/ (2017).
3. Keeling, R. *Scripps Institution of Oceanography* https://scrippsco2.ucsd.edu/ (2017).
4. Grace, J., Mitchard, E. & Gloor, E. Perturbations in the carbon budget of the tropics. *Glob. Chang. Biol.* **20**, 3238–3255 (2014).
5. Schimel, D., Stephens, B. B. & Fisher, J. B. Effect of increasing CO_2 on the terrestrial carbon cycle. *Proc. Natl Acad. Sci. USA* **112**, 436–441 (2015).
This study reconciles atmospheric inversions with models, and so confirms that there is a strong tropical carbon sink, driven by rising atmospheric CO_2.
6. Pan, Y. et al. A large and persistent carbon sink in the world's forests. *Science* **333**, 988–993 (2011).
This paper combines evidence from thousands of global forest plots and other data to estimate the size and location of forest sinks and sources.
7. Patra, P. K. et al. The Orbiting Carbon Observatory (OCO-2) tracks 2–3 peta-gram increase in carbon release to the atmosphere during the 2014–2016 El Niño. *Sci. Rep.* **7**, 13567 (2017).
8. Baccini, A. et al. Tropical forests are a net carbon source based on aboveground measurements of gain and loss. *Science* **358**, 230–234 (2017).
9. Cox, P. M. et al. Sensitivity of tropical carbon to climate change constrained by carbon dioxide variability. *Nature* **494**, 341–344 (2013).
10. Ma, X. et al. Drought rapidly diminishes the large net CO_2 uptake in 2011 over semi-arid Australia. *Sci. Rep.* **6**, 37747 (2016).
11. Liu, J., Bowman, K. W., Schimel, D. S., Parazoo, N. C., Jiang, Z., Lee, M. et al. Contrasting carbon cycle responses of the tropical continents to the 2015–2016 El Niño. *Science* **358**, eaam5690 (2017).
This study uses a satellite sensitive to atmospheric greenhouse gas concentrations to show that all tropical forest areas released CO_2 in response to the 2015–2016 El Niño, but for different reasons.
12. Poorter, L. et al. Biomass resilience of Neotropical secondary forests. *Nature* **530**, 211–214 (2016).
13. Brienen, R. J. W. et al. Long-term decline of the Amazon carbon sink. *Nature* **519**, 344–348 (2015).
Long-term analysis of Amazon plot data shows that the intact forest sink is reducing in size.
14. Qie, L. et al. Long-term carbon sink in Borneo's forests halted by drought and vulnerable to edge effects. *Nat. Commun.* **8**, 1966 (2017); erratum **9**, 342, (2018).
15. Peylin, P. et al. Global atmospheric carbon budget: results from an ensemble of atmospheric CO_2 inversions. *Biogeosciences* **10**, 6699–6720 (2013).
16. Stephens, B. B. et al. Weak northern and strong tropical land carbon uptake from vertical profiles of atmospheric CO_2. *Science* **316**, 1732–1735 (2007).
17. van der Werf, G. R. et al. CO_2 emissions from forest loss. *Nat. Geosci.* **2**, 737–738 (2009); erratum **2**, 829 (2009).
18. Sitch, S. et al. Recent trends and drivers of regional sources and sinks of carbon dioxide. *Biogeosciences* **12**, 653–679 (2015).
19. Ciais, P. et al. in *Climate Change 2013: The Physical Science Basis, Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Stocker, T. F. et al.) (Cambridge Univ. Press, 2013).
20. Hansen, M. C. et al. High-resolution global maps of 21st-century forest cover change. *Science* **342**, 850–853 (2013).
An important dataset is presented and analysed: global deforestation maps at 30-m resolution from 2000–2012.
21. Keenan, R. J. et al. Dynamics of global forest area: Results from the FAO Global Forest Resources Assessment 2015. *For. Ecol. Manage.* **352**, 9–20 (2015).
22. Houghton, R. A., Byers, B. & Nassikas, A. A. A role for tropical forests in stabilizing atmospheric CO_2. *Nat. Clim. Chang.* **5**, 1022–1023 (2015).
23. Schleussner, C.-F. et al. Science and policy characteristics of the Paris Agreement temperature goal. *Nat. Clim. Chang.* **6**, 827–835 (2016).
24. Saatchi, S. S. et al. Benchmark map of forest carbon stocks in tropical regions across three continents. *Proc. Natl Acad. Sci. USA* **108**, 9899–9904 (2011).
25. Baccini, A. et al. Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps. *Nat. Clim. Chang.* **2**, 182–185 (2012).
26. Avitabile, V. et al. An integrated pan-tropical biomass map using multiple reference datasets. *Glob. Chang. Biol.* **22**, 1406–1420 (2016).
27. Beer, C. et al. Terrestrial gross carbon dioxide uptake: global distribution and covariation with climate. *Science* **329**, 834–838 (2010).
28. Malhi, Y. The productivity metabolism and carbon cycle of tropical forest vegetation. *J. Ecol.* **100**, 65–75 (2012).
This paper presents how tropical forests cycle carbon, and the physiological basis of how this will change with climate change.
29. Malhi, Y. The carbon balance of tropical forest regions 1990–2005. *Curr. Opin. Environ. Sustain.* **2**, 237–244 (2010).
30. Smith, P. et al. in *Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Edenhofer, O. et al.) (Cambridge Univ. Press, 2014).
31. Poulter, B. et al. Contribution of semi-arid ecosystems to interannual variability of the global carbon cycle. *Nature* **509**, 600–603 (2014).
32. Jung, M. et al. Global patterns of land-atmosphere fluxes of carbon dioxide latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. *J. Geophys. Res.* **116**, (2011).
33. Zarin, D. J. et al. Can carbon emissions from tropical deforestation drop by 50% in 5 years? *Glob. Chang. Biol.* **22**, 1336–1347 (2016).
34. Achard, F. et al. Determination of tropical deforestation rates and related carbon losses from 1990 to 2010. *Glob. Chang. Biol.* **20**, 2540–2554 (2014).
35. Houghton, R. A. et al. Carbon emissions from land use and land-cover change. *Biogeosciences* **9**, 5125–5142 (2012).
36. Tyukavina, A. et al. Aboveground carbon loss in natural and managed tropical forests from 2000 to 2012. *Environ. Res. Lett.* **10**, 074002 (2015).
37. Liu, Y. Y. et al. Recent reversal in loss of global terrestrial biomass. *Nat. Clim. Chang.* **5**, 470–474 (2015).
38. Harris, N. L. et al. Baseline map of carbon emissions from deforestation in tropical regions. *Science* **336**, 1573–1576 (2012).
39. Kim, D.-H., Sexton, J. O. & Townshend, J. R. Accelerated deforestation in the humid tropics from the 1990s to the 2000s. *Geophys. Res. Lett.* **42**, 3495–3501 (2015).
40. Mitchard, E. T. A. et al. Markedly divergent estimates of Amazon forest carbon density from ground plots and satellites. *Glob. Ecol. Biogeogr.* **23**, 935–946 (2014).
41. de Andrade, R. B. et al. Scenarios in tropical forest degradation: carbon stock trajectories for REDD+. *Carbon Balance Manag.* **12**, 6 (2017).
42. Bustamante, M. M. C. et al. Toward an integrated monitoring framework to assess the effects of tropical forest degradation and recovery on carbon stocks and biodiversity. *Glob. Chang. Biol.* **22**, 92–109 (2016).
43. Brinck, K. et al. High resolution analysis of tropical forest fragmentation and its impact on the global carbon cycle. *Nat. Commun.* **8**, 14855 (2017).
44. Chaplin-Kramer, R. et al. Degradation in carbon stocks near tropical forest edges. *Nat. Commun.* **6**, 10158 (2015).
45. Ryan, C. M., Berry, N. J. & Joshi, N. Quantifying the causes of deforestation and degradation and creating transparent REDD+ baselines: a method and case study from central Mozambique. *Appl. Geogr.* **53**, 45–54 (2014).
46. Berenguer, E. et al. A large-scale field assessment of carbon stocks in human-modified tropical forests. *Glob. Chang. Biol.* **20**, 3713–3726 (2014).
47. Meyer, V. et al. Detecting tropical forest biomass dynamics from repeated airborne lidar measurements. *Biogeosciences* **10**, 5421–5438 (2013).
48. *Global Forest Resources Assessment* (FAO, 2015).
49. Dargie, G. C. et al. Age, extent and carbon storage of the central Congo Basin peatland complex. *Nature* **542**, 86–90 (2017).
50. Draper, F. C. et al. The distribution and amount of carbon in the largest peatland complex in Amazonia. *Environ. Res. Lett.* **9**, 124017 (2014).

© 2018 Springer Nature Limited. All rights reserved.

51. Page, S. Rieley, J. O., Banks, C. J. Global and regional importance of the tropical peatland carbon pool. *Glob. Chang. Biol.* **17**, 798–818 (2011).
52. Moore, S. et al. Deep instability of deforested tropical peatlands revealed by fluvial organic carbon fluxes. *Nature* **493**, 660–663 (2013).
53. Hooijer, A. et al. Current and future CO_2 emissions from drained peatlands in Southeast Asia. *Biogeosciences* **7**, 1505–1514 (2010).
54. Page, S. E. et al. The amount of carbon released from peat and forest fires in Indonesia during 1997. *Nature* **420**, 61–65 (2002).
This study presents evidence that the peat fires in Indonesia during the 1997 El Niño event resulted in a globally important release of carbon.
55. Roucoux, K. H. et al. Threats to intact tropical peatlands and opportunities for their conservation. *Conserv. Biol.* **31**, 1283–1292 (2017).
56. Bloom, A. A. et al. A global wetland methane emissions and uncertainty dataset for atmospheric chemical transport models (WetCHARTs version 1.0). *Geosci. Model Dev.* **10**, 2141–2156 (2017).
57. Lewis, S. L. et al. Increasing carbon storage in intact African tropical forests. *Nature* **457**, 1003–1006 (2009).
58. Phillips, O. L. et al. Changes in the carbon balance of tropical forests: evidence from long-term plots. *Science* **282**, 439–442 (1998).
59. Sheil, D. A critique of permanent plot methods and analysis with examples from Budongo Forest Uganda. *For. Ecol. Manage.* **77**, 11–34 (1995).
60. Wheeler, C. E. et al. Carbon sequestration and biodiversity following 18 years of active tropical forest restoration. *For. Ecol. Manage.* **373**, 44–55 (2016).
61. Grainger, A. Difficulties in tracking the long-term global trend in tropical forest area. *Proc. Natl Acad. Sci. USA* **105**, 818–823 (2008).
62. Romijn, E. et al. Assessing change in national forest monitoring capacities of 99 tropical countries. *For. Ecol. Manage.* **352**, 109–123 (2015).
63. Moutinho, P., Guerra, R. & Azevedo-Ramos, C. Achieving zero deforestation in the Brazilian Amazon: What is missing? *Elementa* **4**, 000125 (2016).
64. Butsic, V., Baumann, M., Shortland, A., Walker, S. & Kuemmerle, T. Conservation and conflict in the Democratic Republic of Congo: the impacts of warfare mining, and protected areas on deforestation. *Biol. Conserv.* **191**, 266–273 (2015).
65. Milodowski, D. T., Mitchard, E. T. A. & Williams, M. Forest loss maps from regional satellite monitoring systematically underestimate deforestation in two rapidly changing parts of the Amazon. *Environ. Res. Lett.* **12**, 094003 (2017).
66. Swamy, L., Drazen, E., Johnson, W. R. & Bukoski, J. J. The future of tropical forests under the United Nations Sustainable Development Goals. *J. Sustain. For.* **37**, 221–256 (2018).
67. Meyfroidt, P. & Lambin, E. F. Forest transition in Vietnam and displacement of deforestation abroad. *Proc. Natl Acad. Sci. USA* **106**, 16139–16144 (2009).
68. Sonter, L. J. et al. Mining drives extensive deforestation in the Brazilian Amazon. *Nat. Commun.* **8**, 1013 (2017).
69. Bonner, M. T. L., Schmidt, S. & Shoo, L. P. A meta-analytical global comparison of aboveground biomass accumulation between tropical secondary forests and monoculture plantations. *For. Ecol. Manage.* **291**, 73–86 (2013).
70. Boysen, L. R. et al. Global and regional effects of land-use change on climate in 21st century simulations with interactive carbon cycle. *Earth Syst. Dyn.* **5**, 309–319 (2014).
71. Bonan, G. B., Doney, S. C. Climate ecosystems, and planetary futures: The challenge to predict life in Earth system models. *Science* **359**, eaam8328 (2018).
An introduction to ESMs, and how they can be improved.
72. Gibbs, H. K. et al. Tropical forests were the primary sources of new agricultural land in the 1980s and 1990s. *Proc. Natl Acad. Sci. USA* **107**, 16732–16737 (2010).
73. Phillips, O. L. et al. Drought sensitivity of the Amazon rainforest. *Science* **323**, 1344–1347 (2009).
74. Brando, P. M. et al. Abrupt increases in Amazonian tree mortality due to drought-fire interactions. *Proc. Natl Acad. Sci. USA* **111**, 6347–6352 (2014).
75. Rowland, L. et al. Death from drought in tropical forests is triggered by hydraulics not carbon starvation. *Nature* **528**, 119–122 (2015).
76. Betts, R. A. et al. Climate and land use change impacts on global terrestrial ecosystems and river flows in the HadGEM2-ES Earth system model using the representative concentration pathways. *Biogeosciences* **12**, 1317–1338 (2015).
77. Ahlström, A., Schurgers, G., Arneth, A. & Smith, B. Robustness and uncertainty in terrestrial ecosystem carbon response to CMIP5 climate change projections. *Environ. Res. Lett.* **7**, 044008 (2012).
78. Stouffer, R. J. et al. CMIP5 scientific gaps and recommendations for CMIP6. *Bull. Am. Meteorol. Soc.* **98**, 95–105 (2017).
79. Moss, R. H. et al. The next generation of scenarios for climate change research and assessment. *Nature* **463**, 747–756 (2010).
80. Ahlström, A., Xia, J., Arneth, A., Luo, Y. & Smith, B. Importance of vegetation dynamics for future terrestrial carbon cycling. *Environ. Res. Lett.* **10**, 054019 (2015).
81. Paris Agreement, Article 5 (UNFCCC, 2015).
82. Turnhout, E. et al. Envisioning REDD+ in a post-Paris era: between evolving expectations and current practice. *WIREs Clim. Chang.* **8**, e425 (2017).
83. Rossi, V. et al. Could REDD+ mechanisms induce logging companies to reduce forest degradation in Central Africa? *J. For. Econ.* **29**, 107–117 (2017).
84. FORESTS: Action Statements and Action Plan (United Nations, 2014)
85. Transforming our world: the 2030 Agenda for Sustainable Development (United Nations, 2015).
86. Veldman, J. W., Silveira, F. A. O., Fleischman, F. D., Ascarrunz, N. L. & Durigan, G. Grassy biomes: an inconvenient reality for large-scale forest restoration? A comment on the essay by Chazdon and Laestadius. *Am. J. Bot.* **104**, 649–651 (2017).
87. Latawiec, A. E., Strassburg, B. B. N., Brancalion, P. H. S., Rodrigues, R. R. & Gardner, T. Creating space for large-scale restoration in tropical agricultural landscapes. *Front. Ecol. Environ.* **13**, 211–218 (2015).
88. Cavaleri, M. A., Reed, S. C., Smith, W. K. & Wood, T. E. Urgent need for warming experiments in tropical forests. *Glob. Chang. Biol.* **21**, 2111–2121 (2015).
89. Chave, J. et al. Improved allometric models to estimate the aboveground biomass of tropical trees. *Glob. Chang. Biol.* **20**, 3177–3190 (2014).
90. Doughty, C. E. et al. What controls variation in carbon use efficiency among Amazonian tropical forests? *Biotropica* **50**, 16–25 (2018).
91. Nepstad, D. C., Tohver, I. M., Ray, D., Moutinho, P. & Cardinot, G. Mortality of large trees and lianas following experimental drought in an Amazon forest. *Ecology* **88**, 2259–2269 (2007).

Acknowledgements The author acknowledges partial support from the Natural Environment Research Council (grant NE/R000751/1) and the UK Space Agency (grant Forests 2020).

Reviewer information *Nature* thanks P. Brando, J. Chave and Y. Malhi for their contribution to the peer review of this work.

Competing interests The authors declare no competing interests.

Additional information
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to E.T.A.M.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

REVIEW

https://doi.org/10.1038/s41586-018-0252-6

El Niño–Southern Oscillation complexity

Axel Timmermann[1,2,3]*, Soon-Il An[4], Jong-Seong Kug[5], Fei-Fei Jin[6], Wenju Cai[7,8,9], Antonietta Capotondi[10,11], Kim Cobb[12], Matthieu Lengaigne[13], Michael J. McPhaden[14], Malte F. Stuecker[15,16], Karl Stein[1,2], Andrew T. Wittenberg[17], Kyung-Sook Yun[1,2], Tobias Bayr[18], Han-Ching Chen[19], Yoshimitsu Chikamoto[20], Boris Dewitte[21,22], Dietmar Dommenget[23], Pamela Grothe[24], Eric Guilyardi[25,26], Yoo-Geun Ham[27], Michiya Hayashi[6], Sarah Ineson[28], Daehyun Kang[29], Sunyong Kim[5], WonMoo Kim[30], June-Yi Lee[1,2], Tim Li[3,6], Jing-Jia Luo[31], Shayne McGregor[23], Yann Planton[25], Scott Power[31], Harun Rashid[7], Hong-Li Ren[32], Agus Santoso[33], Ken Takahashi[34], Alexander Todd[35], Guomin Wang[31], Guojian Wang[7], Ruihuang Xie[36], Woo-Hyun Yang[5], Sang-Wook Yeh[37], Jinho Yoon[38], Elke Zeller[1,2] & Xuebin Zhang[39]

El Niño events are characterized by surface warming of the tropical Pacific Ocean and weakening of equatorial trade winds that occur every few years. Such conditions are accompanied by changes in atmospheric and oceanic circulation, affecting global climate, marine and terrestrial ecosystems, fisheries and human activities. The alternation of warm El Niño and cold La Niña conditions, referred to as the El Niño–Southern Oscillation (ENSO), represents the strongest year-to-year fluctuation of the global climate system. Here we provide a synopsis of our current understanding of the spatio-temporal complexity of this important climate mode and its influence on the Earth system.

The view on the El Niño phenomenon—originally described in 1893 as 'corriente del Niño'[1], a warm regional ocean current that affected climate off the coast of Peru—has changed over the past century. In the 1960s, ENSO was recognized as a basin-scale phenomenon involving coupled atmosphere–ocean processes[2]. A major international research programme in the 1980s and 1990s fundamentally advanced the ability to observe, understand and predict ENSO and its world-wide impacts[3]. During the past 20 years, our understanding of ENSO has continued to evolve, as new layers of complexity (Box 1) have been identified in ENSO dynamics and predictability. The concept of El Niño has developed from one of a canonical progression of phases from onset to maturity and demise[4] (Fig. 1) to one that accounts for its spatio-temporal complexity (Fig. 2) and varying climatic impacts[5–8] (Fig. 3). We have also come to realize that although ENSO primarily manifests itself as a year-to-year climate fluctuation, its dynamics involves a broad range of processes interacting on timescales that range from weeks[9,10] to decades[11]. Here the diversity in patterns, amplitude and temporal evolution of this climate phenomenon will be referred to as ENSO complexity (Box 1).

The most recent El Niño event[12], in 2015/2016, was initiated in boreal spring by a series of westerly wind events (WWEs) (Box 1, Fig. 3e)—a form of tropical weather noise. The associated wind forcing triggered downwelling oceanic Kelvin waves (Box 1, Fig. 1c), thus reducing the upwelling of cold subsurface waters in the eastern Pacific cold tongue (Box 1) and leading to surface warming in the central and eastern Pacific. The positive sea surface temperature anomaly (SSTA) shifted atmospheric convection from the Western Pacific Warm Pool (Box 1) to the central equatorial Pacific, causing a reduction in equatorial trade winds, which in turn intensified surface warming through the positive Bjerknes feedback (Box 1). The seasonally paced termination of the 2015/2016 event (Fig. 3e) was associated with ocean dynamics and the slow discharge of equatorial heat into off-equatorial regions, thus providing a delayed negative feedback (Box 1). The event started to decline in early 2016 and transitioned into a weak La Niña in mid-2016.

In broad terms, this evolution is common to the other strong El Niño events (Fig. 1), in 1982/1983 and 1997/1998 (Fig. 3c). However, no two events are alike—be they strong, moderate or weak (Figs. 2, 3f–m). This diversity arises from the varying roles of noise forcing (Fig. 3c–e) and of positive and negative coupled atmosphere–ocean feedback processes[13] (Box 1) that act to enhance and suppress the growth of SSTAs, respectively. The complexity of ENSO (Box 1), along with internal atmospheric noise, also translates into a diversity of global impacts[7,14]. When the underlying sea surface temperatures (SSTs) change in the equatorial Pacific, there are shifts in atmospheric deep convection, which in turn cause adjustments of the global Walker circulation (Box 1) and generate stationary atmospheric waves[15] that impact the far reaches of our planet. This perturbed global circulation influences weather variability, leading

[1]Center for Climate Physics, Institute for Basic Science (IBS), Busan, South Korea. [2]Pusan National University, Busan, South Korea. [3]International Pacific Research Center, University of Hawaii at Manoa, Honolulu, HI, USA. [4]Department of Atmospheric Sciences, Yonsei University, Seoul, South Korea. [5]Division of Environmental Science and Engineering, Pohang University of Science and Technology (POSTECH), Pohang, South Korea. [6]Department of Atmospheric Science, SOEST, University of Hawaii at Manoa, Honolulu, HI, USA. [7]CSIRO Oceans and Atmosphere, Aspendale, Victoria, Australia. [8]Physical Oceanography Laboratory/CIMST, Ocean University of China and Qingdao National Laboratory for Marine Science and Technology, Qingdao, China. [9]Centre for Southern Hemisphere Oceans Research (CSHOR), CSIRO Oceans and Atmosphere, Hobart, Tasmania, Australia. [10]Cooperative Institute for Research in Environmental Science, University of Colorado, Boulder, CO, USA. [11]Physical Sciences Division, NOAA Earth System Research Laboratory, Boulder, CO, USA. [12]Earth and Atmospheric Sciences, Georgia Tech, Atlanta, GA, USA. [13]LOCEAN/IPSL, Sorbonne Universités/UPMC-CNRS-IRD-MNHN, Paris, France. [14]Pacific Marine Environmental Laboratory/NOAA, Seattle, WA, USA. [15]Department of Atmospheric Sciences, University of Washington, Seattle, WA, USA. [16]Cooperative Programs for the Advancement of Earth System Science, University Corporation for Atmospheric Research, Boulder, CO, USA. [17]Geophysical Fluid Dynamics Laboratory/NOAA, Princeton, NJ, USA. [18]GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany. [19]Department of Atmospheric Sciences, National Taiwan University, Taipei, Taiwan. [20]Department of Plants, Soils, and Climate, Utah State University, Logan, UT, USA. [21]Centro de Estudios Avanzado en Zonas Áridas (CEAZA), Coquimbo, Chile. [22]Laboratoire d'Etudes en Géophysique et Océanographie Spatiale, Toulouse, France. [23]School of Earth, Atmosphere and Environment, Monash University, Clayton, Victoria, Australia. [24]Department of Earth and Environmental Sciences, University of Mary Washington, Fredericksburg, VA, USA. [25]Laboratoire d'Océanographie et du Climat: Expérimentation et Approches Numériques (LOCEAN), IRD/UPMC/CNRS/MNHN, Paris, France. [26]NCAS-Climate, University of Reading, Reading, UK. [27]Department of Oceanography, Chonnam National University, Gwangju, South Korea. [28]Met Office Hadley Centre, Exeter, UK. [29]School of Urban and Environmental Engineering, Ulsan National Institute of Science and Technology, Ulsan, South Korea. [30]Climate Prediction Department, APEC Climate Center, Busan, South Korea. [31]Australian Bureau of Meteorology, Melbourne, Victoria, Australia. [32]Laboratory for Climate Studies, National Climate Center, China Meteorological Administration, Beijing, China. [33]ARC Centre of Excellence for Climate System Science, Faculty of Science, University of New South Wales, Sydney, New South Wales, Australia. [34]Instituto Geofisico del Peru, Lima, Peru. [35]University of Exeter College of Engineering, Mathematics and Physical Sciences, Exeter, UK. [36]Institute of Oceanology, Chinese Academy of Sciences, Qingdao, China. [37]Department of Marine Sciences and Convergent Technology, Hanyang University, Ansan, South Korea. [38]School of Earth Sciences and Environmental Engineering, Gwangju Institute of Science and Technology, Gwangju, South Korea. [39]CSIRO Ocean and Atmosphere, Hobart, Tasmania, Australia. *e-mail: timmermann@pusan.ac.kr

© 2018 Springer Nature Limited. All rights reserved.

Box 1

ENSO glossary

Bjerknes feedback. Positive ENSO feedback along the Equator, in which a weakened (strengthened) equatorial zonal sea surface temperature (SST) gradient weakens (strengthens) trade winds, which in turn further reduce (increase) the zonal SST gradient.

Combination tones/C-mode. Enhanced spectral energy on timescales of 9 months and 15–18 months, generated by the nonlinear modulation of ENSO by the seasonal cycle, and vice versa. C-modes play an important part in the seasonal turnabout of El Niño events.

Eastern Pacific cold tongue. An eastern equatorial Pacific region characterized by wind-driven upwelling of cold subsurface waters (Fig. 1a, b). The cold tongue warms considerably during eastern Pacific El Niño events and cools during La Niña events.

Ekman feedback. Positive (negative) SST anomalies (SSTAs) weaken (strengthen) the equatorial trade winds, reducing (increasing) the upwelling of cold subsurface water in the eastern equatorial Pacific, thus reinforcing the original SSTA.

ENSO complexity. ENSO complexity expands on the concept of ENSO 'pattern diversity' to include also temporal characteristics (from weather, annual cycle, interannual to decadal timescales), dynamics, predictability and global impacts.

ENSO skewness. Amplitude asymmetry of El Niño and La Niña events, which quantifies the fact that El Niño events attain larger amplitudes than La Niña events. Skewness is a clear indication of nonlinearity in the ENSO cycle.

Equatorial Kelvin wave: Eastward-propagating oceanic internal wave that displaces the interface (thermocline) between warm surface waters and cold subsurface waters. Westerly (easterly) equatorial wind anomalies generate downwelling (upwelling) Kelvin waves, which deepen (shoal) the thermocline in the eastern Pacific and reduce (enhance) the efficiency of climatological upwelling.

Multiplicative noise. Interaction between westerly wind events and underlying SST in the western and central Pacific, in which warmer (colder) SST favours more (fewer) westerly wind events; also referred to as state-dependent noise.

Recharge/discharge. Meridional transport of heat into/out of an equatorial band, driven by changes in near-equatorial wind variations. Recharge/discharge processes have a key role in the initiation and termination of El Niño events.

Thermal damping. Typically a negative feedback arising from SST-induced changes in surface radiative and turbulent heat fluxes in the tropical Pacific. It involves tropical clouds, convection and atmospheric boundary layer physics.

Thermocline feedback. Generally positive feedback operating in the eastern equatorial Pacific, in which a warm (cold) equatorial SSTA weakens (strengthens) equatorial trade winds, leading to mean upwelling of anomalously warm (cold) water.

Westerly wind event: Weather systems in the western and central Pacific that are often associated with an abrupt relaxation of the equatorial trade winds, generating downwelling Kelvin waves and an eastward expansion of the Western Pacific Warm Pool.

Western Pacific Warm Pool. Some of the warmest waters in the worlds' oceans occur in the western tropical Pacific, with temperatures exceeding 28 °C (Fig. 1a, b). The Western Pacific Warm Pool's seasonal north–south migrations have an important role in the termination of El Niño events.

Zonal advective feedback: Positive feedback, particularly effective in the central Pacific, in which a positive (negative) equatorial SSTA weakens (strengthens) equatorial trade winds, reducing (enhancing) the oceanic transport of cold waters from the eastern Pacific.

to massive reorganizations of tropical and extratropical temperature and rainfall patterns[16,17] (Fig. 3f–m).

Palaeo-climate reconstructions of the ENSO phenomenon covering the past ~10,000 years also show a wide range of amplitudes[18], thus highlighting the importance of internal climate processes in modulating ENSO complexity on timescales ranging from decades to centuries. In addition, the activity of reconstructed ENSO variability shows an intensification in the late 20th century relative to other pre-industrial periods[18,19], thus raising the general question of whether external forcings could influence the evolution and amplitude of ENSO. How ENSO responds to greenhouse warming is one of the most compelling outstanding questions[20].

Given the societal and environmental relevance of ENSO, it is paramount to improve our understanding of the processes that control the amplitude, timing, duration, predictability and global impacts of ENSO. Here we assess our current understanding of ENSO dynamical processes and their role in controlling complexity of this fundamental climate feature. Against this backdrop, we highlight areas of uncertainty (see 'A unifying framework') as a stimulus for further research.

A conceptual view of ENSO dynamics

Early efforts to elucidate the dynamics of ENSO focused on the average (composite) evolution of El Niño events[21], capturing the typical evolution of ocean and atmosphere conditions from the early-spring initiation of El Niño, to its winter-time peak and transition to La Niña during the subsequent summer (Fig. 1). The enhanced spectral interannual variability of ENSO (Fig. 3a, b) has been explained by invoking positive atmosphere and ocean feedbacks and delayed negative ocean adjustment feedbacks (Box 1), which together can lead to oscillatory dynamics, as encapsulated by a variety of conceptual ENSO models[22]. Here we focus on the ENSO recharge oscillator model[23] which, in its most general form, can be expressed as:

$$\begin{aligned} \mathrm{d}T_\mathrm{e}/\mathrm{d}t &= I_\mathrm{BJ}T_\mathrm{e} + Fh \\ \mathrm{d}h/\mathrm{d}t &= -\varepsilon h - \alpha T_\mathrm{e} \end{aligned} \quad (1)$$

where T_e and h represent the equatorial eastern Pacific surface temperature and zonal mean thermocline depth, respectively, and $\mathrm{d}T_\mathrm{e}/\mathrm{d}t$ and $\mathrm{d}h/\mathrm{d}t$ are the corresponding time derivatives. The Bjerknes stability index, I_BJ, (also referred to as the BJ index or ENSO linear growth rate; Fig. 1j) depends on a number of processes (such as thermocline, zonal advective and Ekman feedbacks) that reinforce SST, and on negative feedbacks from thermal advection by horizontal mean surface currents and thermal damping by net surface heat fluxes[13] (Box 1). In equation (1), ε represents the damping rate of thermocline depth anomalies. The interannual timescale of the ENSO system is mainly determined by F and α, which describe the thermocline feedback (Box 1) and the slow equatorial recharge–discharge process (Box 1) associated with oceanic heat transport, respectively. For a constant I_BJ, the model describes a linear recharge oscillator: starting from neutral conditions with $T_\mathrm{e} \approx 0$ (typically in boreal winter–spring; Figs. 1c, 2) and a charged thermocline state, $h > 0$, an El Niño event can grow (Figs. 1d, e, 2). While eastern equatorial Pacific SSTAs develop, the thermocline feedback, Fh, (Box 1) further intensifies the growth of the SSTAs by upwelling anomalously warm subsurface waters to the surface in the eastern Pacific cold tongue. Moreover, positive eastern Pacific SSTAs $(T_\mathrm{e} > 0)$ cause a weakening of the equatorial trade winds (Fig. 1d, e). The associated wind-stress curl discharges the equatorial heat through Sverdrup transport (Box 1) and ocean boundary processes (Fig. 1f). The resulting drainage of heat in turn weakens the thermocline feedback, and the phase of the ENSO recharge oscillator can transition into a La Niña state (Fig. 1g, h), which is accompanied by recharging of heat through opposite wind-stress curl anomalies (Fig. 1h).

Comparing the linear oscillator solution of equation (1) (constant I_BJ) with the scatter plot of observed equatorial eastern Pacific temperature and zonal mean thermocline depth anomalies (Fig. 2),

© 2018 Springer Nature Limited. All rights reserved.

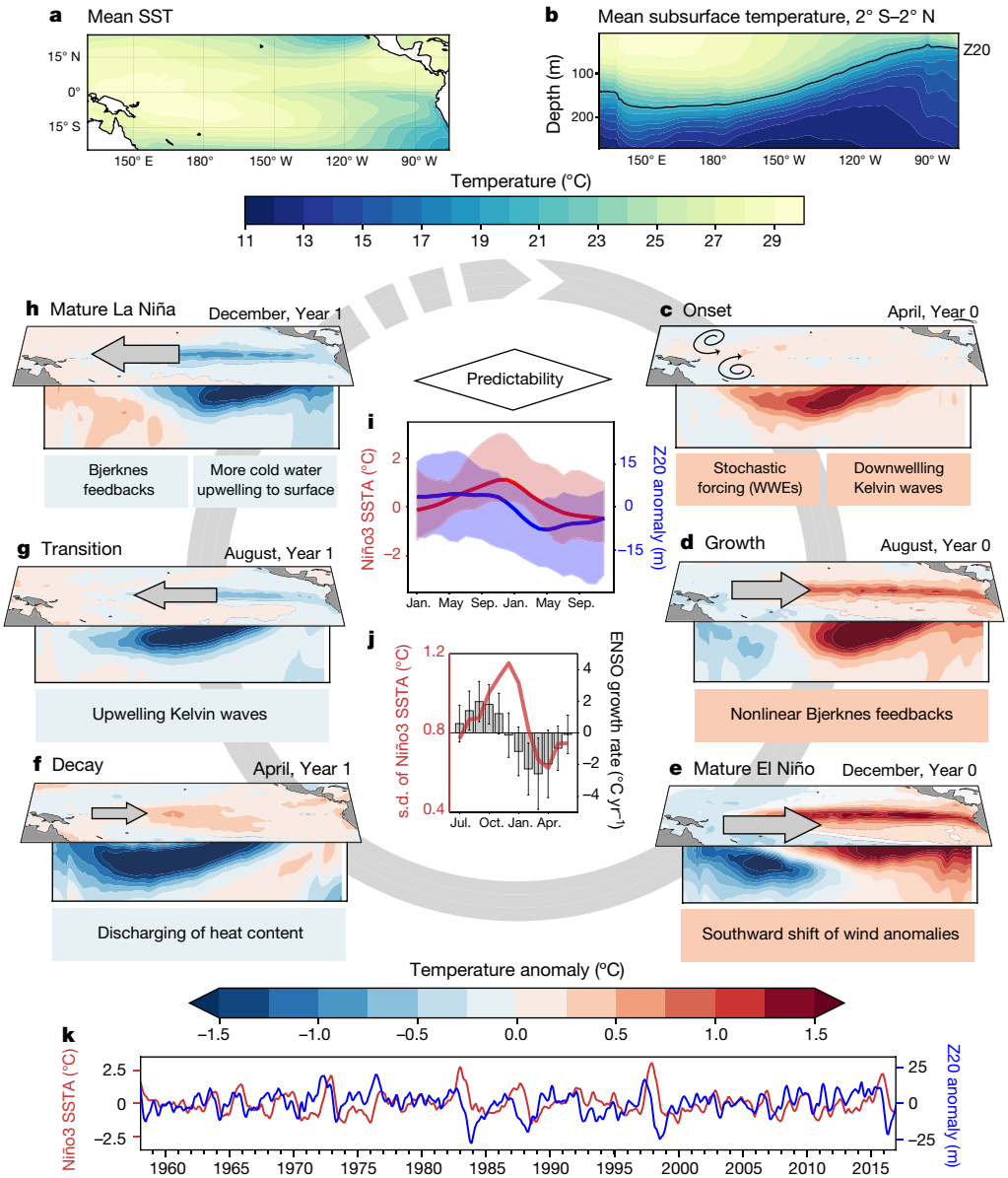


Fig. 1 | ENSO cycle. Composite evolution of El Niño events from 1958 to 2015. **a**, **b**, Mean SST[26] (**a**) and subsurface potential temperature[96] (**b**) between 2° N and 2° S. The depth of the 20 °C isotherm (Z20) is indicated by the black line. **c**–**h**, Composite SSTAs[26] and subsurface temperature anomalies[96] from 17 El Niño events (1963, 1965, 1968, 1969, 1972, 1976, 1977, 1982, 1986, 1987, 1991, 1994, 1997, 2002, 2004, 2006 and 2009), based on the 0.5 °C exceedance of the three-month running mean of the NOAA ERSST.v5 (Extended Reconstructed Sea Surface Temperature dataset, version 5) SSTAs[97] in the Niño3.4 region (averaged over 5° S–5° N and 120° W–170° W). The arrows represent wind anomalies and the boxes list major processes involved in the phases of El Niño evolution. **i**, The seasonal composite means (lines) and spread (shading) of eastern equatorial Pacific SSTA (red; averaged over the Niño3 region: 5° S–5° N and 90° W–150° W) and equatorial Pacific zonal mean Z20 (blue) for the 17 El Niño events. The diamond illustrates that ENSO predictability increases with increasing ENSO signal strength. **j**, The monthly standard deviation (s.d.) of the Niño3 SSTA[26] (red line) and an estimate of the monthly ENSO growth rate based on the Bjerknes stability index. The error bars show the 90% confidence range for the index calculated from the standard error of the regression slope[56]. **k**, Time series of the Niño3 SSTA and zonal mean equatorial Pacific depth anomaly from the 20 °C isotherm (2° S–2°N and 120° E–80° W) from the merged data product[27,98].

we find substantial differences. The observed scatter diagram shows a high degree of irregularity and a notable positive skewness in eastern tropical Pacific SSTAs towards El Niño events (Box 1, Fig. 2). El Niño and La Niña events are very different in terms of their amplitude and time evolution (Figs. 1k, 2a). To account for this additional level of complexity, the simple recharge oscillator model can be extended by including a nonlinear Bjerknes feedback term that represents either atmospheric or oceanic nonlinear processes[24] or multiplicative stochastic forcing[25] (Box 1). For these extensions, the recharge model can then simulate the skewed probability distribution of ENSO (Fig. 2) and the fast growth from neutral to strong El Niño conditions (Fig. 1j). The observed positive skewness of the SSTA (Fig. 2), which indicates the importance of nonlinear dynamical and thermodynamical processes in the coupled tropical Pacific climate system, implies that strong El Niño conditions, which typically last for one year, are on average shorter than La Niña events, which can persist for up to several years (Fig. 1k).

Whereas conceptual models like equation (1) can simulate some key features of ENSO evolution, they can neither explain the presence of its spatial diversity (Fig. 3a, b, f–m) nor the potential remote effects of variability originating from the extra-tropical Pacific, Atlantic or Indian Ocean onto this diversity. An improved framework to characterize and explain ENSO complexity is needed to capture these aspects.

© 2018 Springer Nature Limited. All rights reserved.

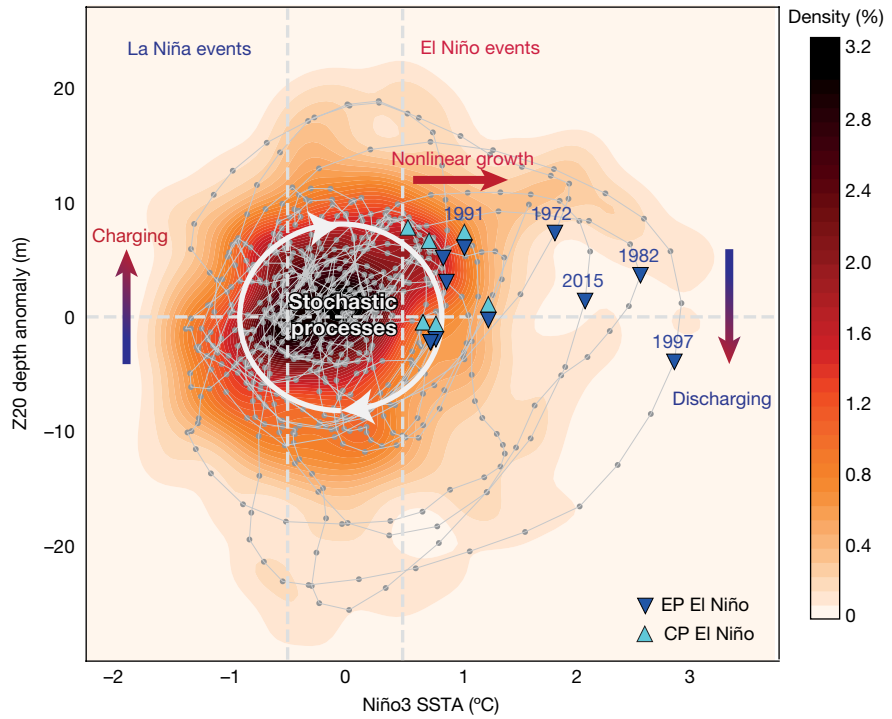


Fig. 2 | Schematic representation of ENSO temporal complexity. Kernel density estimate of the joint probability distribution (orange shading) of the linearly detrended Niño3 SSTA and zonal mean 20 °C isotherm depth anomalies (2° S–2° N and 120° E–80° W) for the period 1958–2016 from the merged data product[27,98]. The grey circles indicate the monthly values of the two time series, smoothed with a three-month running mean filter. Dark and light-blue triangles indicate December values of EP (1972, 1976, 1982, 1986, 1997, 2006 and 2015) and CP (1968, 1994 and 2009) El Niños, respectively. Mixed events (1965, 1991 and 2002) are represented by combined dark and light-blue triangles. The years of the five largest El Niño events are indicated. The white ellipse in the centre corresponds to the progression of the linear recharge oscillator, and arrows on the left (right) indicate charging (discharging) of subsurface warm water in the equatorial Pacific.

Space–time complexity of ENSO

Despite some prominent commonalities discussed in the previous section (Fig. 1), El Niño events differ considerably from each other in terms of magnitude, spatial pattern, temporal evolution and predictability[5–7] (Figs. 2, 3f–m). To characterize the leading modes of equatorial Pacific SST variability and their diverse timescales, we conduct an empirical orthogonal function (EOF) analysis of observed tropical Pacific SSTAs[26] (Fig. 3), which identifies the leading orthogonal patterns of variability. The leading EOF (Fig. 3a), which corresponds to the classical El Niño pattern with eastern tropical Pacific warming, exhibits variability on quasi-quadrennial timescales (3–7 years; spectral density estimate in Fig. 3a). By contrast, the second EOF, which explains only 25% of the variance of the first mode, is characterized by an east–west zonal SST dipole in the tropical Pacific and has enhanced variance on quasi-biennial and decadal timescales (spectral density estimate in Fig. 3b). The interplay of these two EOFs largely captures the spatial diversity of the observed ENSO mode.

Some El Niño events (for example, that of 1997/1998) are characterized by pronounced warming in the eastern Pacific and are referred to as EP El Niño events (Fig. 3c, f), whereas others (for example, that of 2004/2005) show a stronger positive projection on the second EOF mode, which leads to a more pronounced central Pacific warming, and are known as CP El Niño events (Fig. 3d, h–j). More generally, El Niño events can be viewed as the superposition of the two EOF modes, which results in a continuum of ENSO characteristics[27,28] that capture a mix of EP and CP dynamics (for example, the 1991/1992 and 2015/2016 events) (Fig. 2). La Niña events (for example, that of 1999/2000; Fig. 3g), in addition to being weaker than El Niño events, exhibit less diversity in their spatial patterns[6,29], thus clearly pointing to an asymmetry in the underlying dynamical processes for ENSO.

EP El Niño events (for example, that of 1997/1998; Fig. 3c) tend to involve basin-scale equatorial wind anomalies, a strong relaxation of the zonal tilt of the equatorial thermocline (Fig. 1e), a more prominent role for the thermocline feedback (Box 1), large eastward shifts of tropical Pacific convection and strong discharge of heat content (Fig. 3c) away from the equatorial region, which boosts the likelihood of transitioning into a La Niña event[6,30]. By contrast, CP El Niño events (for example, that of 2004/2005; Fig. 3d) tend to involve more local wind feedbacks, a stronger role for the zonal advective feedback (Box 1), little reduction in the zonal tilt of the thermocline, weak shifts of convection, earlier termination, little poleward discharge of ocean heat content (Fig. 3d), a stronger role for thermal damping (Box 1) during the decay phase, a reduced likelihood to transition into La Niña and more susceptibility to disruption by wind noise[6,30]. Compared to CP El Niños, strong EP El Niños also tend to terminate later in boreal spring owing to trade wind collapse, which suppresses the upwelling that normally connects the SST to the evolving thermocline depth[31].

The spatial diversity in the SSTA patterns of ENSO is also associated with different tropical precipitation patterns (Fig. 3f–m), resulting in potentially different remote teleconnection patterns and corresponding weather and climate impacts[7,32]. However, given the high level of internal atmospheric variability[33] and the brevity of the historical record, it has remained difficult to unequivocally detect the differences in the impacts of various ENSO spatial modes. In addition to its spatial diversity, ENSO also exhibits substantial diversity in its temporal evolution (Figs. 1k, 3c–e). Understanding this diversity of El Niño events is crucial for predicting the regional impacts of ENSO, for example, on precipitation patterns, tropical cyclones and other types of severe weather[5]. The extent to which El Niño diversity is predictable relates to whether ENSO complexity originates mainly from random processes or from low-frequency deterministic dynamics. Random processes affecting a single physical ENSO mode could generate diversity in amplitude, spatial structure and temporal evolution[8], consistent with a spatial flavour continuum generated by different realizations of atmospheric noise[27].

© 2018 Springer Nature Limited. All rights reserved.

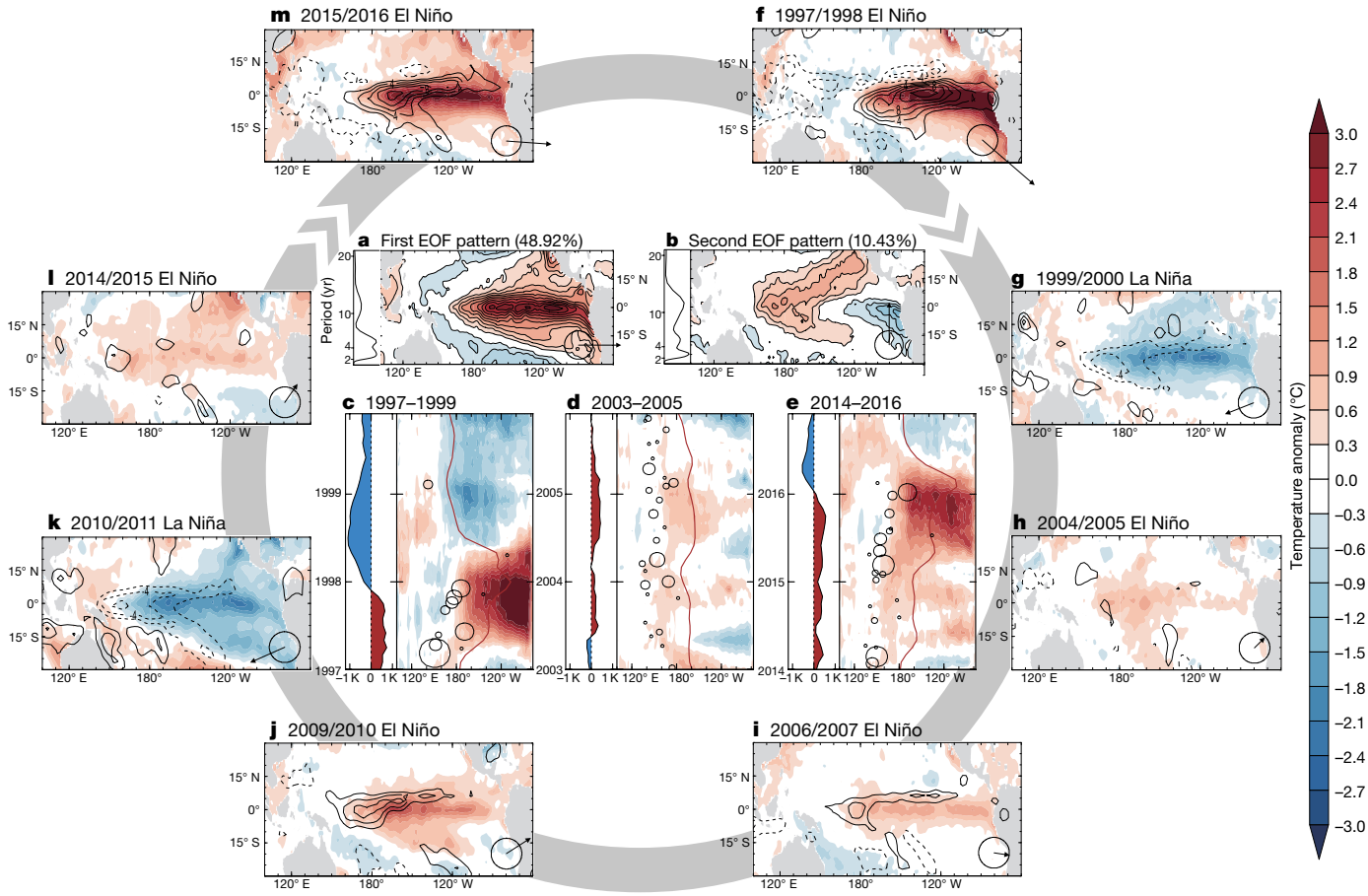


Fig. 3 | Spatio-temporal complexity of ENSO. a, **b**, First (**a**) and second (**b**) EOF patterns of a linearly detrended SSTA[95] computed for 25° S–25° N and 140° E–80° W during 1920–2016, with the associated variance-preserving spectral power density of the normalized principal components shown on the left (the vertical axis is the period in years and the horizontal axis is the logarithm of the power). Numbers in parentheses indicate the variance of the EOF modes. **c**–**e**, Longitude–time evolution of the Pacific SSTA averaged over 5° S–5° N for selected observed ENSO events, the 28.5 °C isotherm of the SST (red curve) representing the edge of the Western Pacific Warm Pool, the longitude and strength of WWEs[99] (black circles) and (on the left) the associated equatorial Pacific heat-content anomaly (temperature anomaly averaged over the top 300 m of the ocean and between 5° S–5° N and 120° E–90° W; range from −1 K to 1 K; red, positive; blue, negative). **f**–**m**, Spatial pattern of SSTA (shaded) and precipitation anomaly[100] (contours; solid line, positive; dashed line, negative; 2 mm d^{-1} interval; zero contour omitted) averaged over the November–January season of selected ENSO events. We note that strong warm events (1997/98 and 2015/16) induce very strong eastward and equatorward shifts of rainfall. At the bottom right of **a**, **b** and **f**–**m** we show the associated principal components (PCs); namely, the projection of each SSTA spatial pattern onto the EOF patterns in **a** and **b**. The abscissa is PC1, the ordinate is PC2 and the arrow length is proportional to the magnitude in PC1–PC2 space (an arrow magnitude of 1 is indicated by the circles).

Alternatively, initial subsurface ocean conditions could modulate the role of stochastic wind forcing in producing diversity. For example, climate model simulations have demonstrated that in the presence of stochastic WWEs, an initial build-up of equatorial Pacific upper-ocean heat content can favour the development of EP, rather than CP, El Niño events[10,34] (Fig. 3d, e). At the onset of strong El Niño events[35], such as those of 1997/1998 and 2015/2016 (Fig. 3c, e), WWE activity tends to strengthen and expand eastwards with the expansion of the Western Pacific Warm Pool and the relaxation of the trade winds. These WWE changes can be parameterized in equation (1) as multiplicative noise (Box 1), which can contribute to ENSO diversity and asymmetries[9,36].

Studies suggest that ENSO diversity may be triggered by climate phenomena outside the tropical Pacific, including the North[37] and South[38] Pacific meridional modes, extra-tropical atmospheric circulation patterns and tropical Atlantic variability[5,39,40]. For example, the negative phase of the North Pacific Oscillation[41] tends to favour the development of positive SSTAs in the central Pacific by weakening the trade winds in the Northern Hemisphere, while the positive phase of the South Pacific Oscillation tends to weaken the Southern Hemisphere trade winds, thereby favouring the development of positive SSTAs in the eastern Pacific. Such remote influences appear to be mediated primarily by how they project onto wind variations in the equatorial Pacific. In essence, westerly wind anomalies in the western equatorial Pacific tend to favour CP El Niños, whereas westerly wind anomalies in the central–eastern equatorial Pacific tend to favour EP El Niños. These external influences can precede the peak of El Niño by 2–3 seasons[39,41] and may provide additional predictability to the spatial characteristics of an emerging El Niño event.

Since 1998, CP events have become more prevalent than EP events[42]. Such a decadal modulation in ENSO diversity is consistent with coupled general circulation models (CGCMs) that can spontaneously generate multidecadal variations in ENSO diversity even in the absence of external radiative forcings[43]. Low-frequency climatic drivers (including natural and anthropogenic forcings)—which involve basin-wide changes in the zonal SST gradient, thermocline depth and winds[44,45]—may also have contributed to the observed decadal swings in ENSO diversity by favouring particular spatio-temporal modes[46]. At this stage, the observational record remains too short to quantify all the possible sources of the decadal modulation of ENSO characteristics.

The current generation of climate models underestimates ENSO diversity[47]. This issue is related to the models' systematic biases, which affect the mean state and ENSO feedbacks. Sources of these biases include deficiencies in the simulation of clouds, atmospheric convection and oceanic mixing[48]. In particular, atmospheric model responses

© 2018 Springer Nature Limited. All rights reserved.

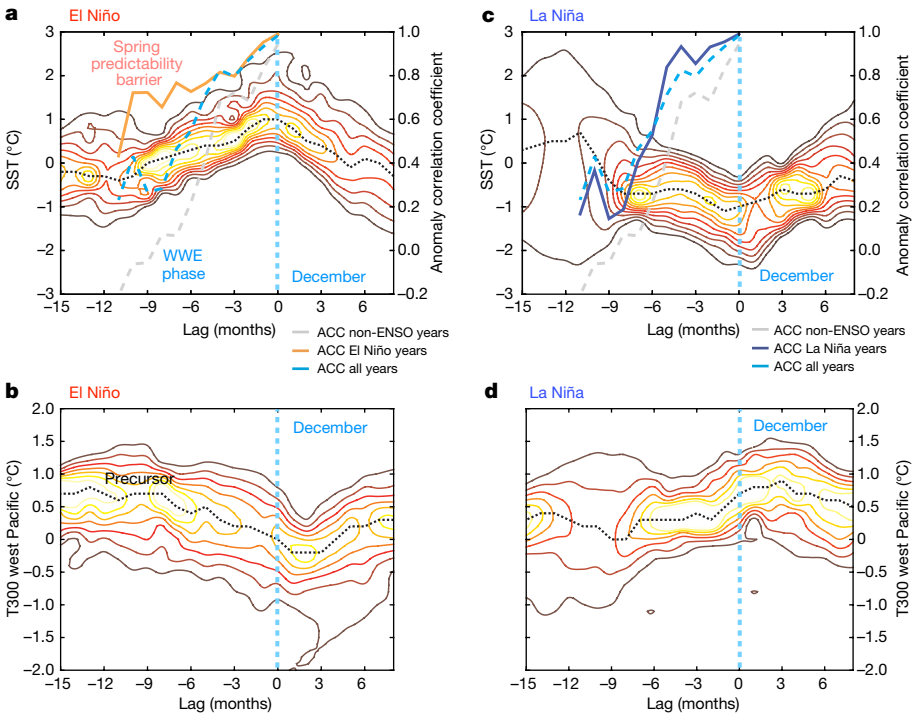


Fig. 4 | Probabilistic ENSO precursors and predictive skill. a, **c**, Time-evolving Kernel density probability density estimates (contours) of the linearly detrended eastern equatorial Pacific SSTA[26] (averaged over the Niño3.4 region: 5° S–5° N and 170° W–120° W) for the period 1958–2016 for El Niño (**a**) and La Niña (**c**) conditions, exceeding the +0.5 °C and −0.5 °C threshold, respectively. **b**, **d**, The same as **a** and **c**, but for western tropical Pacific heat-content anomalies[96]—namely, temperature anomalies averaged over 5° S–5° N, 120° E–155° W and 0–300 m (T300) and high-pass-filtered with a cut-off period of 20 years to remove multi-decadal trends. Every time the probability for SSTAs (heat-content anomalies) to be in the range from −3 °C to 3 °C (−2 °C to 2 °C) is ~100%. However, some SSTA and heat-content values are more likely to occur than others. This probability is indicated by the coloured contours (probability contour interval, 0.005). The time evolution of the probability density estimates is shown for different lead and lag times, relative to El Niño and La Niña events peaking in December. Black dotted lines correspond to the maximum probability for each lag. Thick lines in **a** represent the anomaly correlation coefficient skill (ACC) for the December Niño3.4 SSTA[97] (1980–2015) exceeding +0.5 °C (orange), within the range −0.5 °C to +0.5 °C (grey) and for all years (cyan), calculated using nine coupled models from the North American Multimodel Ensemble project[77]. Lines in **c** are the same as in **a**, but for anomalies below –0.5 °C (dark blue), within the range −0.5 °C to +0.5 °C (grey) and all years (cyan). We note that the ACC shown here in orange and dark blue lines does not represent the skill aggregated over all initial conditions, but only over those identified a posteriori as El Niño and La Niña events.

tend to be relatively insensitive to distinct patterns of SSTAs owing to climatological dry and cold biases in the equatorial central Pacific[47,49].

Seasonal ENSO dynamics

ENSO displays a close relationship with the seasonal cycle[21,50]: El Niño events usually start in boreal spring (Figs. 1c, i, 4), grow during the summer and autumn (Fig. 1d), reach their maximum intensity in winter (Fig. 1e, i) and decay rapidly during late winter and spring (Fig. 1f, j). In most cases, they transition to La Niña events (Figs. 1g, h, 4) by the subsequent summer. This seasonal synchronization of ENSO translates into the observed eastern equatorial Pacific SSTA variance peaking during boreal winter and attaining minimum values during spring (Fig. 1j). It also leads to pronounced seasonal contrasts in the climate impacts and predictability of ENSO (Fig. 1i). ENSO influences the global atmospheric circulation, affecting, for instance, the Asian monsoons[51] and the climate in America[52] and Australia[53].

Randomly occurring sequences of WWEs, typically during spring, can lead to an initial warming of the central eastern equatorial Pacific[54,55] (Fig. 1c). This initial SSTA can grow because the air–sea coupling is strongest in summer and early autumn[56,57] (Fig. 1j). Proposed physical processes for this summer–autumn coupling maximum include (i) the shift of the Intertropical Convergence Zone towards the Equator and its associated increase in western Pacific surface wind convergence[58], (ii) the seasonal outcropping of the equatorial thermocline[59], (iii) the seasonal cooling of the eastern equatorial Pacific[58] and (iv) the reduction of negative cloud feedbacks[60].

The decay of El Niño events typically starts in boreal winter. The anomalous westerly wind anomalies shift southwards from the Equator, leading to a shoaling of the eastern Pacific thermocline and a subsequent reduction of the overlying SSTA[61] (Figs. 1f, 4a). This shift arises from climatological expansion of the Western Pacific Warm Pool into the Southern Hemisphere, coincident with the development of the South Pacific Convergence Zone[62]. In this season, the increased surface heat flux damping[60] results in a decrease of the air–sea coupling strength (Fig. 1j), which—together with the aforementioned seasonal southward wind shift[63] and the equatorial heat content discharge[23] (Fig. 4b)—leads to a rapid transition to a La Niña state.

While these seasonal processes generally operate for different flavours and phases of ENSO, differences in their relative importance can contribute to ENSO complexity. For instance, CP events typically terminate earlier and are less likely to transition to a La Niña state compared to EP El Niño events[62] (Fig. 3d). Furthermore, La Niña conditions can last up to 2–3 years (Figs. 3c, 4b). The ability to simulate ENSO seasonal synchronization for different types of El Niño events varies strongly among the current generation of climate models, probably owing to biases in mean state and seasonal cycle[56].

The influence of the seasonal variations of the air–sea coupling strength discussed above can be included in the framework of the recharge oscillator (equation (1)) by adding a seasonally varying growth rate (I_{BJ}). As expected, this model then captures the observed ENSO seasonal synchronization characteristics, including the seasonal ENSO variance modulation and partial phase synchronization[25]. Interactions between the seasonal cycle in I_{BJ} and the interannual ENSO temperature signal generate variance with periods of roughly 9 and 15–18 months, the so-called combination tone frequencies (Box 1) that broaden the ENSO spectrum predominantly towards higher frequencies[64,65].

© 2018 Springer Nature Limited. All rights reserved.

These interacting dynamics create specific atmospheric circulation patterns that are together referred to as combination mode[64] (C-mode) (Box 1). The spatial pattern of the C-mode exhibits a pronounced hemispheric asymmetry, which includes an anomalous cyclonic low-level wind circulation in the southern central tropical Pacific and an anomalous anticyclonic low-level wind circulation in the northwestern tropical Pacific. Some of the prominent local expressions are the aforementioned southward shift of equatorial wind anomalies[63] (Fig. 1e) and the anomalous western North Pacific anticyclone[65].

ENSO predictability

To link our understanding of dynamical tropical air–sea interactions with ENSO predictability, it is helpful to elucidate the seasonal evolution of (i) potential precursors that may contribute to long-term predictability[66] (9–15 months lead time), (ii) triggers that can rapidly increase the likelihood for event development (6–9 months lead time) and (iii) transition processes (see 'A conceptual view of ENSO dynamics' and 'Seasonal ENSO dynamics'). The development of a typical EP event can be divided into different seasonal stages, which each contribute differently to the boreal winter 6–9 month SSTA forecasting skill in the Niño 3.4 region (central to eastern Pacific), as illustrated by the anomaly correlation coefficient skill between seasonal forecasts performed with the North American Multimodel Ensemble[67] (NMME) and the observations (Fig. 4a, cyan dashed line). Prior to boreal spring, a charged western tropical Pacific heat content is a necessary condition for the subsequent development of El Niño events (Fig. 4b). Thus, the corresponding warm pool heat advection processes[68,69] have a key role in determining the long-term memory for ENSO forecasts. Furthermore, atmospheric precursors in the North[37] and South Pacific[38], the Indian Ocean[70] or the tropical Atlantic[40,71] have been suggested to influence the El Niño evolution for long lead times.

It must be emphasized here that the presence of such early oceanic or atmospheric precursors is usually not sufficient for El Niño growth because one of the key trigger mechanisms is the stochastic WWE activity in boreal spring and early summer[72]. This is clearly illustrated by the fact that even though initial heat-content conditions were favourable for El Niño development in early 2012, 2014 and 2017, the subsequent SSTA growth stayed below expectations. Individual WWEs are not predictable beyond the weather prediction horizon, which implies that, on average, forecasts initialized in boreal spring have relatively low long-term skill[73], in particular in the absence of precursor signals (Fig. 4a, cyan line). However, precursor signals in western tropical Pacific heat content (Fig. 4b) could be indicative of potentially developing El Niño conditions, which in effect enhance the predictability (Fig. 4a, orange line) relative to the averaged case (Fig. 4a, cyan line). The competing roles of stochasticity versus ocean memory for this so-called spring predictability barrier (Fig. 4a) and for long-lead-time forecasts have been intensely debated[74,75].

If a sufficient amount of westerly momentum is transferred in boreal spring from the atmosphere to the ocean, zonal advective processes begin moving the warm pool front eastwards, and downwelling Kelvin waves (Box 1) generate surface warming in the eastern tropical Pacific about two months later. These anomalies are further intensified (Fig. 4a) owing to increasing summer air–sea coupling strength (Fig. 1j), while anomalously warm water is drained from the Western Pacific Warm Pool (Fig. 4b). This phase exhibits a high degree of climate predictability, as documented by the high anomaly correlation coefficients (>0.6) between predicted boreal winter El Niño events and observations for forecasts initialized in boreal summer (Fig. 4a). The subsequent demise of an El Niño event is largely controlled by ocean subsurface processes and the discharge of zonal heat content away from the Equator (Figs. 1i, 2, equation (1)), as well as by the seasonally modulated southward shift of westerly wind anomalies[63], which in turn leads to a relaxation of the zonally integrated thermocline anomalies. This seasonally locked decay of El Niño conditions under a low-noise atmospheric environment further contributes to the long-term averaged ENSO prediction skill[62].

The subsequent evolution into a La Niña state (Fig. 4a–c) and the possibility of having multi-year La Niña events (Figs. 1k, 4c) are less well understood. La Niña events are often preceded by a strong El Niño. However, as indicated by the broad probability distribution of the SSTA at a lag of 9–15 months (Fig. 4c), other initial conditions can also develop into La Niña events peaking in boreal winter (Fig. 4c). During the transition from El Niño to La Niña, equatorial heat gets quickly discharged, and 6–9 months before a peak La Niña in boreal winter, we observe the smallest values of the equatorial heat content (Fig. 1i) and a slow recharging tendency of the Western Pacific Warm Pool (Fig. 4d). However, during this period the probability density of western tropical Pacific heat-content anomalies is relatively broad, which translates into an overall reduction of predictive skill (Fig. 4c). Because longer-lasting La Niña events are exposed to a variety of atmospheric and oceanic perturbations and the annual cycle, a dynamical decoupling of La Niña and subsequent El Niño events may occur[76]. In boreal winter, during the peak of the La Niña, the Western Pacific Warm Pool is fully charged (Fig. 4d) to values that are typical for an El Niño precursor (Fig. 4b). However, the SST conditions do not necessarily have to swing back to an El Niño state and sometimes even a second La Niña can develop. When comparing the anomaly correlation coefficient skill for December La Niña target conditions with the averaged skill for all years (1980–2015) from the NMME[77], we find very little difference (Fig. 4c) for lead times of 1–12 months, which suggests that (i) La Niña conditions have a considerably lower predictability than El Niño, (ii) the predictability of La Niña is to a first order captured well by the mean statistical skill of the current generation of seasonal prediction models. Using ensemble forecasting techniques, a recent study[78] identified potential predictors for the likelihood of multi-year La Niña events, which include the magnitude of thermocline discharge and the amplitude of the preceding El Niño event, suggesting the possibility for longer-term forecasts also for La Niña.

How the different stages of predictability differ between CP and EP events and whether there are distinct precursor patterns for different ENSO flavours still remains controversial[68,79]. Despite an improved understanding of ENSO dynamics, the ENSO prediction skill has not demonstrated a steady improvement during the past few decades, with even a decrease at the turn of the 21st century[73]. This decrease may be related to the reduced ENSO amplitude and the more frequent occurrence of CP events during that period[79], as their evolution and climate impacts tend to be less predictable than those of EP El Niño events[80].

A unifying framework

The previous discussion has highlighted a variety of dynamical pathways that can be combined to explain the spatio-temporal complexity of the ENSO phenomenon (Fig. 5). Extending beyond the simple single-mode theory (equation (1); 'A conceptual view of ENSO dynamics'), which captures several—but not all—features of ENSO dynamics, our proposed framework for ENSO complexity is based on the co-existence of a duplet of linear eigenmodes (Fig. 5a, b), which can be derived from a deterministic, intermediate-complexity tropical atmosphere and ocean model[46] and a number of excitation mechanisms. These two generic coupled eigenmodes are characterized by spatial patterns that closely resemble the observed EP and CP modes (Fig. 5) and by timescales of approximately four and two years, respectively. The four-year (quasi-quadrennial) mode is more prominent (Fig. 5, lower left) when the mean thermocline is deep and the trade winds are weak, and it relies strongly on thermocline feedback. By contrast, the two-year (quasi-biennial) mode is dominant when the mean thermocline is shallow and the equatorial trade winds are strong. Its SST variability is strongly controlled by zonal advective feedback[46]. These features are akin to their observational counterparts (Figs. 3, 5c, d). For realistic background states, both modes operate not far from criticality (zero growth rate) (Fig. 5, lower left), which implies that they can be easily excited by other processes. Their stability and excitability depends further on the prevailing climatic background conditions.

© 2018 Springer Nature Limited. All rights reserved.

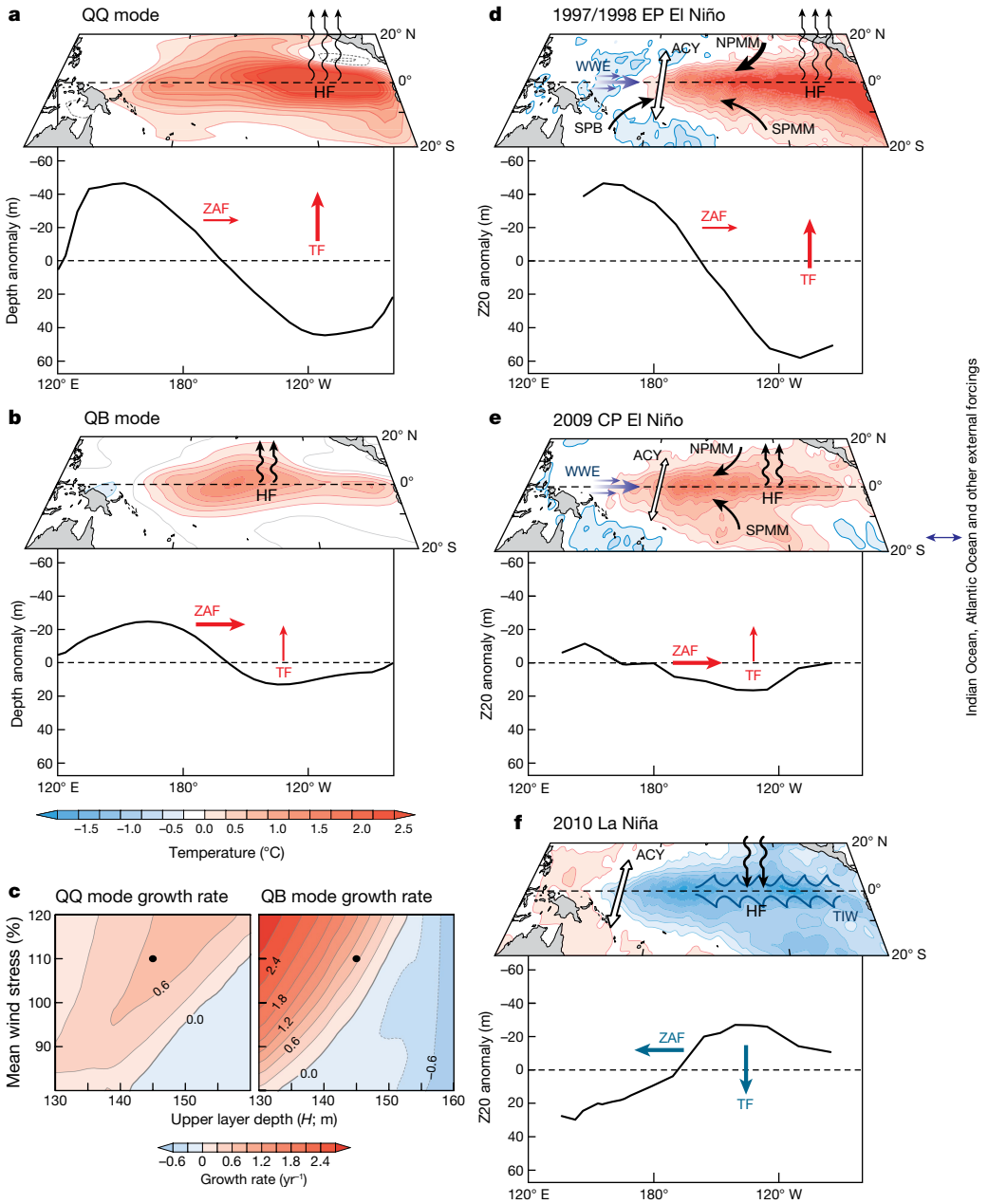


Fig. 5 | Mechanisms of ENSO complexity. a, **b**, Leading two eigenmodes of tropical Pacific SSTA and equatorial thermocline depth anomalies (averaged between 5° S–5° N) with periods of about 4 yr (QQ, quasi-quadrennial) and about 2 yr (QB, quasi-biennial), calculated fom an intermediate ENSO model[46]. The differences in zonal location of the centre in SSTAs and thermocline anomalies are largely due to the different roles of the zonal advective feedback (ZAF) and thermocline feedback (TF). **c**, Growth rates of the two eigenmodes as a function of the mean thermocline depth, *H*, and the mean strength of equatorial trade winds relative to climatological conditions. Black dots mark the mean state for the modes displayed in **a** and **b**. **d**–**f**, Patterns of SSTAs[26] and equatorial Tropical Atmosphere Ocean (TAO)/ Triangle Trans-Ocean Buoy Network (TRITON) 20 °C thermocline depth anomalies for typical EP (1997/1998) and CP (2009/2010) El Niño and La Niña (boreal winter 2010) events (November–January), with schematic representations of the key excitation, nonlinear and cross-scale interaction mechanisms: annual cycle (ACY), WWEs, South Pacific booster (SPB), North and South Pacific meridional modes (NPMM and SPMM, respectively) and tropical instabilty waves (TIW). The solid red, eastward (blue, westward) arrows represent the ZAF and the red, upward (blue, downward) arrows denote the TF for El Niño (La Niña) conditions. The relative sizes and different zonal positions of the arrows indicate qualitatively the strength and areas of strong feedback efficiency. Curly upward (downward) arrows denote damping net surface heat flux (HF) feedback for El Niño (La Niña).

At the heart of our explanation for the spatial flavours of ENSO is the aforementioned multiplicity of coupled ENSO eigenmodes (Fig. 5a, b), as seen in an intermediate ENSO model[46]. Furthermore, the temporal complexity is generated in part by the different oscillation frequencies of the quasi-quadrennial and quasi-biennial modes and additionally by different external excitation processes. Such processes are associated, for example, with the North and South Pacific meridional modes[81,82], the South Pacific booster[38], WWEs (see 'Space–time complexity of ENSO' and 'Seasonal ENSO dynamics'), tropical instability waves[83] or transbasin influences[40] (Fig. 5). In particular, asymmetric dependencies related to the increased WWE activity during El Niño and enhanced tropical instability wave activity during La Niña make these cross-scale interactions very effective sources for ENSO complexity. Furthermore, the annual cycle of winds and SSTs has a key role in determining the seasonal timing of ENSO anomalies and its predictability (see 'Seasonal ENSO dynamics' and 'ENSO predictability'). To further explain the fact that El Niño anomalies are stronger in amplitude (see 'A conceptual view of ENSO dynamics') and exhibit a more pronounced spatial

© 2018 Springer Nature Limited. All rights reserved.

diversity (see 'Space–time complexity of ENSO') and higher predictability[84] relative to their La Niña counterparts (Figs. 4b, d, 5c, e), we need to invoke additional nonlinear processes. Nonlinearities, particularly in atmospheric deep convection and oceanic heat advection, can induce a wide range of additional timescales[64] and new spatial structures[85,86] by potentially coupling or amplifying the duplet of ENSO eigenmodes.

Decadal subsurface processes[87,88] can affect the long-term climatological background state. In turn, this background state changes the stability of the two primary ENSO eigenmodes (Fig. 5a, b) and their excitability. Hence, slow background state changes in the Pacific Ocean can have a key role in generating and modulating the spatio-temporal complexity of ENSO.

Our unifying framework for ENSO complexity (Fig. 5), which identifies key factors for ENSO complexity (primary ENSO eigenmodes, excitation processes, nonlinearities and cross-timescale interactions), may serve as a roadmap for further hypothesis testing, process studies and diagnostic analysis of climate models. It can also help guide the evolution of the tropical Pacific observing system, which is essential for underpinning ENSO research and forecasting[89]. In addition, this framework can be used to determine how the shortcomings in representing ENSO complexity in climate and Earth system models are related to a variety of feedback processes and biases in the mean state and annual cycle that affect the generation of climate variability.

Outlook

The reliability of dynamical seasonal climate predictions depends heavily on the representation of ENSO processes in CGCMs and on the continuous improvement of the global ocean observing system. Climate models still exhibit stubborn climate biases in the eastern equatorial Pacific[90] that may affect their representation of feedbacks (see 'A conceptual view of ENSO dynamics') and ENSO complexity[47], as well as the fidelity of operational ENSO forecasts. Identifying and resolving underlying systematic model biases will help in developing the next generation of models for seamless climate forecasts and projections.

Future research on ENSO complexity needs to address the role of the seasonal cycle for CP ENSO dynamics, the near-absence of spatial diversity for La Niña[29] (Fig. 5e), the impact of decadal background-state changes on ENSO modes vis-à-vis multiple-timescale processes involving WWEs, tropical instability waves, extratropical triggers, as well as the response of the spatio-temporal complexity of ENSO to past and future climate change. Moreover, whether the underlying dynamical origin of spatio-temporal diversity in CGCMs can in fact be linked to the duplet of quasi-quadrennial and quasi-biennial ENSO eigenmodes must be investigated. This can be tested by applying interactive atmosphere ensemble averaging techniques in coupled climate models[91], which artificially reduce non-SST-related atmospheric perturbations. Moreover, the use of flux-adjusted CGCMs[92] could help elucidate how model biases impact the spatial diversity of ENSO and provide a more effective way of improving seasonal climate predictions. Such experiments could reveal if there are distinct precursors for ENSO diversity, which could be used to further inform ENSO forecasts. Much scientific emphasis has been placed on understanding the growth of El Niño events. However, given the severe impacts of La Niña—for example, on drought in the southwestern United States[93] or the Horn of Africa[94]—and the fact that La Niña events may last longer than one year (Fig. 4), it will be paramount to gain also deeper understanding of the processes controlling La Niña and its predictability through observational, diagnostic and modelling studies.

A growing global population in the 21st century has become increasingly vulnerable to natural hazards as human activities alter the climate and the environment. Society therefore has an urgent demand for better climate products and services, including improved ENSO monitoring and predictions and long-term projections, to better inform decision making for agriculture and food security, public health, water resource management, energy production, human migration and disaster risk reduction. Because ENSO involves a broad range of disciplines, from atmosphere and ocean dynamics, to ecosystems and societal impacts, it is a unifying concept in Earth system science[95]. Thus, our proposed framework for ENSO complexity can serve as both a catalyst to further research and, in its practical applications, an essential contributor for sustainable development and environmental stewardship in a changing world.

Received: 25 October 2017; Accepted: 2 March 2018;
Published online 25 July 2018.

1. Carrillo, C. N. Hidrografía oceánica. *Bol. Soc. Geogr. Lima* **1**, 72–110 (1893).
2. Bjerknes, J. Atmospheric teleconnections from the equatorial Pacific. *Mon. Weath. Rev.* **97**, 163–172 (1969).
3. McPhaden, M. J., Busalacchi, A. J. & Anderson, D. L. T. A TOGA retrospective. *Oceanography* **23**, 86–103 (2010).
4. Cai, W. J. et al. More extreme swings of the South Pacific convergence zone due to greenhouse warming. *Nature* **488**, 365–369 (2012).
5. Capotondi, A. et al. Understanding ENSO diversity. *Bull. Am. Meteorol. Soc.* **96**, 921–938 (2015).
6. Kug, J. S., Jin, F. F. & An, S. I. Two types of El Niño events: cold tongue El Niño and warm pool El Niño. *J. Clim.* **22**, 1499–1515 (2009).
This study demonstrates that the two types of El Niño (CP and EP) have different dynamical structures, including discharge processes and dominant SST feedbacks.
7. Ashok, K., Behera, S. K., Rao, S. A., Weng, H. Y. & Yamagata, T. El Niño Modoki and its possible teleconnection. *J. Geophys. Res. Oceans* **112**, C11007 (2007).
8. Takahashi, K., Montecinos, A., Goubanova, K. & Dewitte, B. ENSO regimes: reinterpreting the canonical and Modoki El Niño. *Geophys. Res. Lett.* **38**, L10704 (2011).
9. Tziperman, E. & Yu, L. S. Quantifying the dependence of westerly wind bursts on the large-scale tropical Pacific SST. *J. Clim.* **20**, 2760–2768 (2007).
10. Lengaigne, M. et al. Triggering of El Niño by westerly wind events in a coupled general circulation model. *Clim. Dyn.* **23**, 601–620 (2004).
11. Choi, J., An, S. I. & Yeh, S. W. Decadal amplitude modulation of two types of ENSO and its relationship with the mean state. *Clim. Dyn.* **38**, 2631–2644 (2012).
12. L'Heureux, M. L. et al. Observing and predicting the 2015/16 El Niño. *Bull. Am. Meteorol. Soc.* **98**, 1363–1382 (2017).
13. Jin, F. F., Kim, S. T. & Bejarano, L. A coupled-stability index for ENSO. *Geophys. Res. Lett.* **33**, L23708 (2006).
14. Hoerling, M. P. & Kumar, A. Why do North American climate anomalies differ from one El Niño event to another? *Geophys. Res. Lett.* **24**, 1059–1062 (1997).
15. Karoly, D. J. & Hoskins, B. J. 3 dimensional propagation of planetary waves. *J. Meteorol. Soc. Jpn* **60**, 109–123 (1982).
16. Ropelewski, C. F. & Halpert, M. S. Global and regional scale precipitation patterns associated with the El-Niño Southern Oscillation. *Mon. Weath. Rev.* **115**, 1606–1626 (1987).
17. Larkin, N. K. & Harrison, D. E. On the definition of El Niño and associated seasonal average US weather anomalies. *Geophys. Res. Lett.* **32**, L13705 (2005).
18. Cobb, K. M. et al. Highly variable El Niño-Southern Oscillation throughout the holocene. *Science* **339**, 67–70 (2013).
19. McGregor, S., Timmermann, A., England, M. H., Timm, O. E. & Wittenberg, A. T. Inferred changes in El Niño–Southern Oscillation variance over the past six centuries. *Clim. Past* **9**, 2269–2284 (2013).
This study uses multi-proxy data to demonstrate that ENSO variance has increased over the past century relative to the previous 500 years.
20. Cai, W. J. et al. ENSO and greenhouse warming. *Nat. Clim. Change* **5**, 849–859 (2015).
21. Rasmusson, E. M. & Carpenter, T. H. Variations in tropical sea-surface temperature and surface wind fields associated with the Southern Oscillation El-Niño. *Mon. Weath. Rev.* **110**, 354–384 (1982).
22. Wang, C. Z. On the ENSO mechanisms. *Adv. Atmos. Sci.* **18**, 674–691 (2001).
23. Jin, F. F. An equatorial ocean recharge paradigm for ENSO. Part I: conceptual model. *J. Atmos. Sci.* **54**, 811–829 (1997).
This paper introduces a heuristic model that explains key features of ENSO dynamics, such as the important role of recharge and discharge processes in ENSO.
24. Roberts, A., Guckenheimer, J., Widiasih, E., Timmermann, A. & Jones, C. Mixed-mode oscillations of El Niño-Southern Oscillation. *J. Atmos. Sci.* **73**, 1755–1766 (2016).
25. Levine, A. F. Z. & Jin, F. F. Noise-induced instability in the ENSO recharge oscillator. *J. Atmos. Sci.* **67**, 529–542 (2010).
26. Rayner, N. A. et al. Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res. Atmos.* **108**, 4407 (2003).
27. Giese, B. S. & Ray, S. El Niño variability in simple ocean data assimilation (SODA), 1871–2008. *J. Geophys. Res. Oceans* **116**, (2011).
28. Johnson, N. C. How many ENSO flavors can we distinguish? *J. Clim.* **26**, 4816–4827 (2013).
29. Kug, J. S. & Ham, Y. G. Are there two types of La Niña? *Geophys. Res. Lett.* **38**, L16704 (2011).
30. Kao, H. Y. & Yu, J. Y. Contrasting eastern-Pacific and central-Pacific types of ENSO. *J. Clim.* **22**, 615–632 (2009).
31. Lengaigne, M. & Vecchi, G. A. Contrasting the termination of moderate and extreme El Niño events in coupled general circulation models. *Clim. Dyn.* **35**, 299–313 (2010).

© 2018 Springer Nature Limited. All rights reserved.

32. An, S.-I., Kug, J.-S., Timmermann, A., Kang, I.-S. & Timm, O. The influence of ENSO on the generation of decadal variability in the North Pacific. *J. Clim.* **20**, 667–680 (2007).
33. Deser, C., Simpson, I. R., McKinnon, K. A. & Phillips, A. S. The Northern Hemisphere extratropical atmospheric circulation response to ENSO: how well do we know it and how do we evaluate models accordingly? *J. Clim.* **30**, 5059–5082 (2017).
34. Fedorov, A. V., Hu, S. N., Lengaigne, M. & Guilyardi, E. The impact of westerly wind bursts and ocean initial state on the development, and diversity of El Niño events. *Clim. Dyn.* **44**, 1381–1401 (2015).
35. Seiki, A. & Takayabu, Y. N. Westerly wind bursts and their relationship with intraseasonal variations and ENSO. Part I: statistics. *Mon. Weath. Rev.* **135**, 3325–3345 (2007).
36. Jin, F. F., Lin, L., Timmermann, A. & Zhao, J. Ensemble-mean dynamics of the ENSO recharge oscillator under state-dependent stochastic forcing. *Geophys. Res. Lett.* **34**, L03807 (2007).
37. Vimont, D. J., Battisti, D. S. & Hirst, A. C. The seasonal footprinting mechanism in the CSIRO general circulation models. *J. Clim.* **16**, 2653–2667 (2003).
This article presents a coupled air-sea mechanism by which extratropical Pacific climate anomalies can propagate into the tropics.
38. Hong, L. C. L Ho & Jin, F. F. A Southern Hemisphere booster of super El Niño. *Geophys. Res. Lett.* **41**, 2142–2149 (2014).
39. Ham, Y. G., Kug, J. S., Park, J. Y. & Jin, F. F. Sea surface temperature in the north tropical Atlantic as a trigger for El Niño/Southern Oscillation events. *Nat. Geosci.* **6**, 112–116 (2013).
40. Chikamoto, Y. et al. Skilful multi-year predictions of tropical trans-basin climate variability. *Nat. Commun.* **6**, 6869 (2015).
This article documents the impact of Atlantic SSTAs on the generation of CP El Niño events.
41. Yu, J. Y. & Kim, S. T. Relationships between extratropical sea level pressure variations and the central Pacific and eastern Pacific types of ENSO. *J. Clim.* **24**, 708–720 (2011).
42. Lee, T. & McPhaden, M. J. Increasing intensity of El Niño in the central-equatorial Pacific. *Geophys. Res. Lett.* **37**, L14603 (2010).
43. Wittenberg, A. T., Rosati, A., Delworth, T. L., Vecchi, G. A. & Zeng, F. R. ENSO modulation: is it decadally predictable? *J. Clim.* **27**, 2667–2681 (2014).
44. Yeh, S. W. et al. El Niño in a changing climate. *Nature* **461**, 511–514 (2009); erratum 462, 674 (2009).
45. Capotondi, A. & Sardeshmukh, P. D. Optimal precursors of different types of ENSO events. *Geophys. Res. Lett.* **42**, 9952–9960 (2015).
46. Xie, R. & Jin, F.-F. Two leading ENSO modes and El Niño Types in the Zebiak–Cane Model. *J. Clim.* (2018).
This study provides theoretical evidence for two coupled modes that resemble EP and CP El Niño events with different timescales and background-state sensitivities.
47. Ham, Y. G. & Kug, J. S. How well do current climate models simulate two types of El Niño? *Clim. Dyn.* **39**, 383–398 (2012).
This work shows that current climate models tend to underestimate the diversity of El Niño owing to dry and cold biases in the equatorial central Pacific.
48. Bellenger, H., Guilyardi, E., Leloup, J., Lengaigne, M. & Vialard, J. ENSO representation in climate models: from CMIP3 to CMIP5. *Clim. Dyn.* **42**, 1999–2018 (2014).
49. Bayr, T. et al. Mean-state dependence of ENSO atmospheric feedbacks in climate models. *Clim. Dyn.* **50**, 3171–3194 (2017).
50. Li, T. Phase transition of the El Niño Southern Oscillation: a stationary SST mode. *J. Atmos. Sci.* **54**, 2872–2887 (1997).
This study applied seasonally varying instability to model coupled ENSO dynamics.
51. Zhang, W. J. et al. Unraveling El Niño's impact on the East Asian monsoon and Yangtze River summer flooding. *Geophys. Res. Lett.* **43**, 11375–11382 (2016).
52. Trenberth, K. E. et al. Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperatures. *J. Geophys. Res. Oceans* **103**, 14291–14324 (1998).
53. Risbey, J. S., Pook, M. J., McIntosh, P. C., Wheeler, M. C. & Hendon, H. H. On the remote drivers of rainfall variability in Australia. *Mon. Weath. Rev.* **137**, 3233–3253 (2009).
54. McPhaden, M. J. Genesis and evolution of the 1997–98 El Niño. *Science* **283**, 950–954 (1999).
55. Vecchi, G. A. & Harrison, D. E. Tropical Pacific sea surface temperature anomalies, El Niño, and equatorial westerly wind events. *J. Clim.* **13**, 1814–1830 (2000).
56. Wengel, C., Latif, M., Park, W., Harlass, J. & Bayr, T. Seasonal ENSO phase locking in the Kiel Climate Model: the importance of the equatorial cold sea surface temperature bias. *Clim. Dyn.* **50**, 901–919 (2017).
57. Zebiak, S. E. & Cane, M. A. A model El-Niño Southern Oscillation. *Mon. Weath. Rev.* **115**, 2262–2278 (1987).
58. Tziperman, E., Zebiak, S. E. & Cane, M. A. Mechanisms of seasonal–ENSO interaction. *J. Atmos. Sci.* **54**, 61–71 (1997).
59. Galanti, E. et al. The equatorial thermocline outcropping—a seasonal control on the tropical Pacific Ocean–atmosphere instability strength. *J. Clim.* **15**, 2721–2739 (2002).
60. Dommenget, D. & Yu, Y. S. The seasonally changing cloud feedbacks contribution to the ENSO seasonal phase-locking. *Clim. Dyn.* **47**, 3661–3672 (2016).
61. Harrison, D. E. & Vecchi, G. A. On the termination of El Niño. *Geophys. Res. Lett.* **26**, 1593–1596 (1999).
62. Lengaigne, M., Boulanger, J. P., Menkes, C. & Spencer, H. Influence of the seasonal cycle on the termination of El Niño events in a coupled general circulation model. *J. Clim.* **19**, 1850–1868 (2006).
63. McGregor, S., Timmermann, A., Schneider, N., Stuecker, M. F. & England, M. H. The effect of the south Pacific convergence zone on the termination of El Niño events and the meridional asymmetry of ENSO. *J. Clim.* **25**, 5566–5586 (2012).
This work demonstrates that the southward wind shift that leads to the rapid decay of El Niño events is related to the seasonal formation of the South Pacific Convergence Zone.
64. Stuecker, M. F., Timmermann, A., Jin, F. F., McGregor, S. & Ren, H. L. A combination mode of the annual cycle and the El Niño/Southern Oscillation. *Nat. Geosci.* **6**, 540–544 (2013).
65. Stuecker, M. F., Jin, F. F., Timmermann, A. & McGregor, S. Combination mode dynamics of the anomalous northwest Pacific anticyclone. *J. Clim.* **28**, 1093–1111 (2015).
66. Meinen, C. S. & McPhaden, M. J. Observations of warm water volume changes in the equatorial Pacific and their relationship to El Niño and La Niña. *J. Clim.* **13**, 3551–3559 (2000).
67. Barnston, A., Tippett, M. K., Ranganathan, M. & L'Heureux, M. Deterministic skill of ENSO predictions from the North American Multimodel Ensemble. *Clim. Dyn.* https://doi.org/10.1007/s00382-017-3603-3 (2017); erratum https://doi.org/10.1007/s00382-017-3814-7 (2017)
68. Ramesh, N. & Murtugudde, R. All flavours of El Niño have similar early subsurface origins. *Nat. Clim. Change* **3**, 42–46 (2013).
69. Ballester, J., Bordoni, S., Petrova, D. & Rodo, X. Heat advection processes leading to El Niño events as depicted by an ensemble of ocean assimilation products. *J. Geophys. Res. Oceans* **121**, 3710–3729 (2016).
70. Izumo, T. et al. Influence of the state of the Indian Ocean Dipole on the following year's El Niño. *Nat. Geosci.* **3**, 168–172 (2010).
71. Ham, Y. G., Kug, J. S. & Park, J. Y. Two distinct roles of Atlantic SSTs in ENSO variability: North Tropical Atlantic SST and Atlantic Niño. *Geophys. Res. Lett.* **40**, 4012–4017 (2013).
72. Takahashi, K. & Dewitte, B. Strong and moderate nonlinear El Niño regimes. *Clim. Dyn.* **46**, 1627–1645 (2016).
73. Barnston, A. G., Tippett, M. K., L'Heureux, M. L., Li, S. H. & DeWitt, D. G. Skill of real-time seasonal ENSO model predictions during 2002–11: is our capability increasing? *Bull. Am. Meteorol. Soc.* **93**, 631–651 (2012).
74. Newman, M. & Sardeshmukh, P. D. Are we near the predictability limit of tropical Indo-Pacific sea surface temperatures? *Geophys. Res. Lett.* **44**, 8520–8529 (2017).
75. Petrova, D., Koopman, S. J., Ballester, J. & Rodo, X. Improving the long-lead predictability of El Niño using a novel forecasting scheme based on a dynamic components model. *Clim. Dyn.* **48**, 1249–1276 (2017).
76. Kessler, W. S. Is ENSO a cycle or a series of events? *Geophys. Res. Lett.* **29**, 2125 (2002).
This study challenges the notion of ENSO as a cycle and highlights the fact that the ENSO system can lose its dynamical memory during long La Niña events.
77. Barnston, A. G. & Tippett, M. K. Do statistical pattern corrections improve seasonal climate predictions in the North American Multimodel Ensemble models? *J. Clim.* **30**, 8335–8355 (2017).
78. DiNezio, P. N., Deser, C., Okumura, Y. & Karspeck, A. Predictability of 2-year La Niña events in a coupled general circulation model. *Clim. Dyn.* **49**, 4237–4261 (2017).
79. McPhaden, M. J. A. 21st century shift in the relationship between ENSO SST and warm water volume anomalies. *Geophys. Res. Lett.* **39**, L09706 (2012).
80. Jeong, H. I. et al. Assessment of the APCC coupled MME suite in predicting the distinctive climate impacts of two flavors of ENSO during boreal winter. *Clim. Dyn.* **39**, 475–493 (2012).
81. Larson, S. & Kirtman, B. The Pacific Meridional Mode as a trigger for ENSO in a high-resolution coupled model. *Geophys. Res. Lett.* **40**, 3189–3194 (2013).
82. Zhang, H. H., Clement, A. & Di Nezio, P. The south Pacific meridional mode: a mechanism for ENSO-like variability. *J. Clim.* **27**, 769–783 (2014).
83. An, S. I. Interannual variations of the Tropical Ocean instability wave and ENSO. *J. Clim.* **21**, 3680–3686 (2008).
84. Larson, S. M. & Kirtman, B. P. Linking preconditioning to extreme ENSO events and reduced ensemble spread. *Clim. Dyn.* https://doi.org/10.1007/s00382-017-3791-x (2017).
85. Chen, N. & Majda, A. J. Simple dynamical models capturing the key features of the Central Pacific El Niño. *Proc. Natl Acad. Sci. USA* **113**, 11732–11737 (2016).
86. Hao, Z., Neelin, J. D. & Jin, F. F. Nonlinear tropical air-sea interaction in the fast-wave limit. *J. Clim.* **6**, 1523–1544 (1993).
87. Schneider, N. The response of tropical climate to the equatorial emergence of spiciness anomalies. *J. Clim.* **17**, 1083–1095 (2004).
88. McGregor, S., Sen Gupta, A., Holbrook, N. J. & Power, S. B. The modulation of ENSO variability in CCSM3 by extratropical Rossby waves. *J. Clim.* **22**, 5839–5853 (2009).
89. Xue, Y. et al. A real-time ocean reanalyses intercomparison project in the context of tropical pacific observing system and ENSO monitoring. *Clim. Dyn.* **49**, 3647–3672 (2017).
90. Widlansky, M. J. et al. Changes in South Pacific rainfall bands in a warming climate. *Nat. Clim. Change* **3**, 417–423 (2013).
91. Kirtman, B. P., Straus, D. M., Min, D. H., Schneider, E. K. & Siqueira, L. Toward linking weather and climate in the interactive ensemble NCAR climate model. *Geophys. Res. Lett.* **36**, (2009).

© 2018 Springer Nature Limited. All rights reserved.

92. Vecchi, G. A. et al. On the seasonal forecasting of regional tropical cyclone activity. *J. Clim.* **27**, 7994–8016 (2014).
93. Cook, E. R., Seager, R., Cane, M. A. & Stahle, D. W. North American drought: reconstructions, causes, and consequences. *Earth Sci. Rev.* **81**, 93–134 (2007).
94. Nicholson, S. E. & Selato, J. C. The influence of La Niña on African rainfall. *Int. J. Climatol.* **20**, 1761–1776 (2000).
95. McPhaden, M. J., Zebiak, S. E. & Glantz, M. H. ENSO as an integrating concept in Earth science. *Science* **314**, 1740–1745 (2006).
96. Balmaseda, M. A., Mogensen, K. & Weaver, A. T. Evaluation of the ECMWF ocean reanalysis system ORAS4. *Q. J. R. Meteorol. Soc.* **139**, 1132–1161 (2013).
97. Huang, B. Y. et al. Extended reconstructed sea surface temperature, version 5 (ERSSTv5): upgrades, validations, and intercomparisons. *J. Clim.* **30**, 8179–8205 (2017).
98. Penny, S. G., Behringer, D. W., Carton, J. A. & Kalnay, E. A hybrid global ocean data assimilation system at NCEP. *Mon. Weath. Rev.* **143**, 4660–4677 (2015).
99. Puy, M., Vialard, J., Lengaigne, M. & Guilyardi, E. Modulation of equatorial Pacific westerly/easterly wind events by the Madden–Julian oscillation and convectively-coupled Rossby waves. *Clim. Dyn.* **46**, 2155–2178 (2016).
100. Adler, R. F. et al. The version-2 global precipitation climatology project (GPCP) monthly precipitation analysis (1979–present). *J. Hydrometeorol.* **4**, 1147–1167 (2003).

Acknowledgements A.T., K.S., K.-S.Y. and E.Z. were supported by the Institute for Basic Science (project code IBS-R028-D1). B.D. was funded by Fondecyt (grant 1151185). S.-I.A. was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF-2017R1A2A2A05069383). J.-S.K. was supported by the National Research Foundation of Korea (NRF-2017R1A2B3011511). F.-F.J.'s contribution was sponsored through the US NSF Grant AGS-1406601 and the US Department of Energy Grant DE-SC0005110. T.B. receives funding from SFB 754, project 'Climate–Biochemistry Interactions in the tropical Ocean'. M.J.M. is supported by the US National Oceanic and Atmospheric Administration (NOAA). H.-L.R. is supported by the China Meteorological Special Research Project (grant number GYHY201506013). S.I. was supported by the UK–China Research & Innovation Partnership Fund through the Met Office Climate Science for Service Partnership (CSSP) China as part of the Newton Fund. M.F.S. acknowledges support from the NOAA Climate and Global Change Postdoctoral Fellowship Program, administered by UCAR's Cooperative Programs for the Advancement of Earth System Sciences (CPAESS). H.R. was partly funded by the National Environmental Science Program, Australia. This is PMEL contribution number 4723. The authors thank the TAO Project Office of NOAA/PMEL for providing the TAO/TRITON 20 °C isotherm depth anomaly data shown in Fig. 5.

Reviewer information *Nature* thanks M. L'Heureux, X. Rodo and A. Tudhope for their contribution to the peer review of this work.

Author contributions The manuscript was written as a group effort during the 'El Niño Complexity workshop', held at Pusan National University from 16 to 20 October 2017. All authors contributed to the manuscript preparation and the discussions that led to the final figure selection. A.T., J.-S.K., S.-I.A. and F.-F.J. designed the study and served as coordinating lead authors for the sections 'ENSO predictability', 'Space–time complexity of ENSO', 'A conceptual view of ENSO dynamics' and 'A unifying framework', respectively. A.T. oversaw the writing of each section, preparation of figures and selection of references. W.C., A.C. K.C., M.L., M.J.M, M.F.S and A.T.W. served as coordinating lead authors for various sections.

Competing interests The authors declare no competing interests.

Additional information
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to A.T.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

REVIEW

https://doi.org/10.1038/s41586-018-0337-2

Machine learning for molecular and materials science

Keith T. Butler[1], Daniel W. Davies[2], Hugh Cartwright[3], Olexandr Isayev[4]* & Aron Walsh[5,6]*

Here we summarize recent progress in machine learning for the chemical sciences. We outline machine-learning techniques that are suitable for addressing research questions in this domain, as well as future directions for the field. We envisage a future in which the design, synthesis, characterization and application of molecules and materials is accelerated by artificial intelligence.

The Schrödinger equation provides a powerful structure–property relationship for molecules and materials. For a given spatial arrangement of chemical elements, the distribution of electrons and a wide range of physical responses can be described. The development of quantum mechanics provided a rigorous theoretical foundation for the chemical bond. In 1929, Paul Dirac famously proclaimed that the underlying physical laws for the whole of chemistry are "completely known"[1]. John Pople, realizing the importance of rapidly developing computer technologies, created a program—Gaussian 70—that could perform ab initio calculations: predicting the behaviour, for molecules of modest size, purely from the fundamental laws of physics[2]. In the 1960s, the Quantum Chemistry Program Exchange brought quantum chemistry to the masses in the form of useful practical tools[3]. Suddenly, experimentalists with little or no theoretical training could perform quantum calculations too. Using modern algorithms and supercomputers, systems containing thousands of interacting ions and electrons can now be described using approximations to the physical laws that govern the world on the atomic scale[4–6].

The field of computational chemistry has become increasingly predictive in the twenty-first century, with activity in applications as wide ranging as catalyst development for greenhouse gas conversion, materials discovery for energy harvesting and storage, and computer-assisted drug design[7]. The modern chemical-simulation toolkit allows the properties of a compound to be anticipated (with reasonable accuracy) before it has been made in the laboratory. High-throughput computational screening has become routine, giving scientists the ability to calculate the properties of thousands of compounds as part of a single study. In particular, density functional theory (DFT)[8,9], now a mature technique for calculating the structure and behaviour of solids[10], has enabled the development of extensive databases that cover the calculated properties of known and hypothetical systems, including organic and inorganic crystals, single molecules and metal alloys[11–13].

The emergence of contemporary artificial-intelligence methods has the potential to substantially alter and enhance the role of computers in science and engineering. The combination of big data and artificial intelligence has been referred to as both the "fourth paradigm of science"[14] and the "fourth industrial revolution"[15], and the number of applications in the chemical domain is growing at an astounding rate. A subfield of artificial intelligence that has evolved rapidly in recent years is machine learning. At the heart of machine-learning applications lie statistical algorithms whose performance, much like that of a researcher, improves with training. There is a growing infrastructure of machine-learning tools for generating, testing and refining scientific models. Such techniques are suitable for addressing complex problems that involve massive combinatorial spaces or nonlinear processes, which conventional procedures either cannot solve or can tackle only at great computational cost.

As the machinery for artificial intelligence and machine learning matures, important advances are being made not only by those in mainstream artificial-intelligence research, but also by experts in other fields (domain experts) who adopt these approaches for their own purposes. As we detail in Box 1, the resources and tools that facilitate the application of machine-learning techniques mean that the barrier to entry is lower than ever.

In the rest of this Review, we discuss progress in the application of machine learning to address challenges in molecular and materials research. We review the basics of machine-learning approaches, identify areas in which existing methods have the potential to accelerate research and consider the developments that are required to enable more wide-ranging impacts.

Nuts and bolts of machine learning

With machine learning, given enough data and a rule-discovery algorithm, a computer has the ability to determine all known physical laws (and potentially those that are currently unknown) without human input. In traditional computational approaches, the computer is little more than a calculator, employing a hard-coded algorithm provided by a human expert. By contrast, machine-learning approaches learn the rules that underlie a dataset by assessing a portion of that data and building a model to make predictions. We consider the basic steps involved in the construction of a model, as illustrated in Fig. 1; this constitutes a blueprint of the generic workflow that is required for the successful application of machine learning in a materials-discovery process.

Data collection

Machine learning comprises models that learn from existing (training) data. Data may require initial preprocessing, during which missing or spurious elements are identified and handled. For example, the Inorganic Crystal Structure Database (ICSD) currently contains more than 190,000 entries, which have been checked for technical mistakes but are still subject to human and measurement errors. Identifying and removing such errors is essential to avoid machine-learning algorithms being misled. There is a growing public concern about the lack of reproducibility and error propagation of experimental data

[1]ISIS Facility, Rutherford Appleton Laboratory, Harwell Campus, Harwell, UK. [2]Department of Chemistry, University of Bath, Bath, UK. [3]Department of Chemistry, Oxford University, Oxford, UK. [4]Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [5]Department of Materials Science and Engineering, Yonsei University, Seoul, South Korea. [6]Department of Materials, Imperial College London, London, UK. *e-mail: olexandr@olexandrisayev.com; a.walsh@imperial.ac.uk

© 2018 Springer Nature Limited. All rights reserved.

Box 1

Learning to learn

One of the most exciting aspects of machine-learning techniques is their potential to democratize molecular and materials modelling by reducing the computer power and prior knowledge required for entry. Just as Pople's Gaussian software made quantum chemistry more accessible to a generation of experimental chemists, machine-learning approaches, if developed and implemented correctly, can broaden the routine application of computer models by non-specialists. The accessibility of machine-learning technology relies on three factors: open data, open software and open education. There is an increasing drive for open data within the physical sciences, with an ideal best practice outlined recently[98,99]. Some of the open software being developed is listed in Table 2. There are also many excellent open education resources available, such as massive open online courses (MOOCs).

fast.ai (http://www.fast.ai) is a course that is "making neural nets uncool again" by making them accessible to a wider community of researchers. One of the advantages of this course is that users start to build working machine-learning models almost immediately. However, it is not for absolute beginners, requiring a working knowledge of computer programming and high-school-level mathematics.

DataCamp (https://www.datacamp.com) offers an excellent introduction to coding for data-driven science and covers many practical analysis tools relevant to chemical datasets. This course features interactive environments for developing and testing code and is suitable for non-coders because it teaches Python at the same time as machine learning.

Academic MOOCs are useful courses for those wishing to get more involved with the theory and principles of artificial intelligence and machine learning, as well as the practice. The Stanford MOOC (https://www.coursera.org/learn/machine-learning) is popular, with excellent alternatives available from sources such as https://www.edx.org (see, for example, 'Learning from data (introductory machine learning)') and https://www.udemy.com (search for 'Machine learning A–Z'). The underlying mathematics is the topic of a course from Imperial College London (https://www.coursera.org/specializations/mathematics-machine-learning).

Many machine-learning professionals run informative blogs and podcasts that deal with specific aspects of machine-learning practice. These are useful resources for general interest as well as for broadening and deepening knowledge. There are too many to provide an exhaustive list here, but we recommend https://machinelearningmastery.com and http://lineardigressions.com as a starting point.

published in peer-reviewed scientific literature. In certain fields, such as cheminformatics, best practices and guidelines have been established to address these problems[16].

The training of a machine-learning model may be supervised, semi-supervised or unsupervised, depending on the type and amount of available data. In supervised learning, the training data consist of sets of input and associated output values. The goal of the algorithm is to derive a function that, given a specific set of input values, predicts the output values to an acceptable degree of fidelity. If the available dataset consists of only input values, unsupervised learning can be used in an attempt to identify trends, patterns or clustering in the data. Semi-supervised learning may be of value if there is a large amount of input data, but only a limited amount of corresponding output data.

Supervised learning is the most mature and powerful of these approaches, and is used in the majority of machine-learning studies in the physical sciences, such as in the mapping of chemical composition to a property of interest. Unsupervised learning is less common, but can be used for more general analysis and classification of data or to identify previously unrecognized patterns in large datasets[17].

Data representation

Even though raw scientific data are usually numerical, the form in which data are presented often affects learning. In many types of spectroscopy, the signal is acquired in the time domain, but for interpretation it is converted to the frequency domain using the Fourier transform. Like scientists, a machine-learning algorithm might learn more effectively using one format than the other. The process of converting raw data into something more suitable for an algorithm is called featurization or feature engineering.

The more suitable the representation of the input data, the more accurately can an algorithm map it to the output data. Selecting how best to represent the data could require insight into both the underlying scientific problem and the operation of the learning algorithm, because it is not always obvious which choice of representation will give the best performance; this is an active topic of research for chemical systems[18].

Many representations are available to encode structures and properties. One example is the Coulomb matrix[19], which contains information on atomic nuclear repulsion and the potential energy of free atoms, and is invariant to molecular translations and rotation. Molecular systems also lend themselves to descriptions as graphs[20]. In the solid state, the conventional description of crystal structures that uses translation vectors and fractional coordinates of the atoms is not appropriate for machine learning because a lattice can be represented in an infinite number of ways by choosing a different coordinate system. Representations based on radial distribution functions[21], Voronoi tessellations[22] or property-labelled materials fragments[23] are among the new ways in which this problem is being tackled.

Choice of learner

When the dataset has been collected and represented appropriately, it is time to choose a model to learn from it. A wide range of model types (or learners) exists for model building and prediction. Supervised-learning models may predict output values within a discrete set (such as categorizing a material as a metal or an insulator) or a continuous set (such as polarizability). Building a model for the former requires classification, whereas the latter requires regression. A range of learning algorithms can be applied (see Table 1), depending on the type of data and the question posed. It may be helpful to use an ensemble of different algorithms, or of similar algorithms with different values for their internal parameters (known as 'bagging' or 'stacking'), to create a more robust overall model. We outline some of the common algorithms (learners) in the following.

Naive Bayes classifiers[24] are a collection of classification algorithms based on Bayes' theorem that identify the most probable hypothesis, given the data as prior knowledge about the problem. Bayes' theorem provides a formal way of calculating the probability that a hypothesis is correct, given a set of existing data. New hypotheses can then be tested and the prior knowledge updated. In this way, the hypothesis (or model) with the highest probability of correctly representing the data can be selected.

In *k*-nearest-neighbour[25] methods, the distances between samples and training data in a descriptor hyperspace are calculated. They are so called because the output value for a prediction relies on the values of the *k* 'nearest neighbours' in the data, where *k* is an integer. Nearest-neighbour models can be used in both classification and regression models: in classification, the prediction is determined by the class of the majority of the *k* nearest points; in regression, it is determined by the average of the *k* nearest points.

Decision trees[26] are flowchart-like diagrams used to determine a course of action or outcome. Each branch of the tree represents a possible decision, occurrence or reaction. The tree is structured to show how and why one choice may lead to the next, with branches indicating that each option is mutually exclusive. Decision trees comprise a root node, leaf nodes and branches. The root node is the starting point of

© 2018 Springer Nature Limited. All rights reserved.

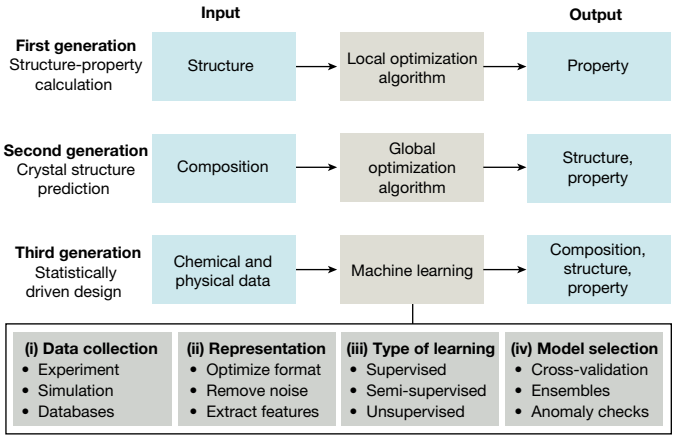


Fig. 1 | Evolution of the research workflow in computational chemistry. The standard paradigm in the first-generation approach is to calculate the physical properties of an input structure, which is often performed via an approximation to the Schrödinger equation combined with local optimization of the atomic forces. In the second-generation approach, by using global optimization (for example, an evolutionary algorithm) an input of chemical composition is mapped to an output that contains predictions of the structure or ensemble of structures that the combination of elements are likely to adopt. The emerging third-generation approach is to use machine-learning techniques with the ability to predict composition, structure and properties provided that sufficient data are available and an appropriate model is trained. Four stages of training a machine-learning model with some of the common choices are listed in the bottom panel.

the tree. Both root and leaf nodes contain questions or criteria to be addressed. Branches are arrows connecting nodes, showing the flow from question to answer. Decision trees are often used in ensemble methods (meta-algorithms), which combine multiple trees into one predictive model to improve performance.

Kernel methods are a class of algorithms, the best known members of which are support vector machine and kernel ridge regression[27]. The name 'kernel' comes from the use of a kernel function—a function that transforms input data into a higher-dimensional representation that makes the problem easier to solve. In a sense, a kernel is a similarity function provided by the domain expert: it takes two inputs and creates an output that quantifies how similar they are.

Artificial neural networks and deep neural networks[28] loosely mimic the operation of the brain, with artificial neurons (the processing unit) arranged in input, output and hidden layers. In the hidden layers, each neuron receives input signals from other neurons, integrates those signals and then uses the result in a straightforward computation. Connections between neurons have weights, the values of which represent the stored knowledge of the network. Learning is the process of adjusting the weights so that the training data are reproduced as accurately as possible.

Whatever the model, most learners are not fully autonomous, requiring at least some guidance. The values of internal variables (hyperparameters) are estimated beforehand using systematic and random searches, or heuristics. Even modest changes in the values of hyperparameters can improve or impair learning considerably, and the selection of optimal values is often problematic. Consequently, the development of automatic optimization algorithms is an area of active investigation, as is their incorporation into accessible packages for non-expert users (see Table 2).

Model optimization

When the learner (or set of learners) has been chosen and predictions are being made, a trial model must be evaluated to allow for optimization and ultimate selection of the best model. Three principal sources of error arise and must be taken into account: model bias, model variance and irreducible errors, with the total error being the sum of these. Bias is the error from incorrect assumptions in the algorithm and can result in the model missing underlying relationships. Variance is sensitivity to small fluctuations in the training set. Even well-trained machine-learning models may contain errors due to noise in the training data, measurement limitations, calculation uncertainties, or simply outliers or missing data. Poor model performance usually indicates a high bias or a high variance, as illustrated in Fig. 2.

High bias (also known as underfitting) occurs when the model is not flexible enough to adequately describe the relationship between inputs and predicted outputs, or when the data are insufficiently detailed to allow the discovery of suitable rules. High variance (or overfitting) occurs when a model becomes too complex; typically, this occurs as the number of parameters is increased. The diagnostic test for overfitting is that the accuracy of a model in representing training data continues to improve, while the performance in estimating test data plateaus or declines.

The key test for the accuracy of a machine-learning model is its successful application to unseen data. A widely used method for determining the quality of a model involves withholding a randomly selected portion of data during training. This withheld dataset, known as a test set, is shown to the model once training is complete (Fig. 2). The extent to which the output data in the validation set is accurately predicted then provides a measure of the effectiveness of training. Cross-validation is reliable only when the samples used for training and validation are representative of the whole population, which may present problems if the sample size is small or if the model is applied to data from compounds that are very different to those in the original dataset. A careful selection of methods for evaluating the transferability and applicability of a model is required in such cases.

Accelerating the scientific method

Whether through the enumeration and analysis of experimental data or the codification of chemical intuition, the application of informatics to guide laboratory chemists is advancing rapidly. In this section, we explore how machine learning is helping to progress and to reduce the barriers between chemical and materials design, synthesis, characterization and modelling. We also describe some of the important developments in the field of artificial intelligence for data-mining existing literature.

Guiding chemical synthesis

Organic chemists were among the first scientists to recognize the potential of computational methods in laboratory practice. E. J. Corey's Organic Chemical Simulation of Synthesis (OCSS) program[29], developed 50 years ago, was an attempt to automate retrosynthetic analysis. In a synthetic chemistry route, the number of possible transformations

Table 1 | Classes of machine-learning techniques and some chemical questions they could answer

Class	Bayesian	Evolutionary[a]	Symbolist	Connectionist	Analogist
Method	Probabilistic inference	Evolving structures	Logical inference	Pattern recognition	Constrained optimization
Algorithms include	Naive Bayes Bayesian networks	Genetic algorithm Particle swarm	Rules Decision trees	Artificial neural networks Back propagation	Nearest neighbour Support vectors
Chemical query	Is my new theory valid?	What molecule gives this property?	How do I make this material?	What compound did I synthesize?	Find a structure–property relation

The classes shown were chosen following ref. [97].
[a]Although evolutionary algorithms are often integrated into machine-learning procedures, they form part of a wider class of stochastic search algorithms.

© 2018 Springer Nature Limited. All rights reserved.

Table 2 | Publicly accessible learning resources and tools related to machine learning

Name	Description	URL
General-purpose machine-learning frameworks		
Caret	Package for machine learning in R	https://topepo.github.io/caret
Deeplearning4j	Distributed deep learning for Java	https://deeplearning4j.org
H2O.ai	Machine-learning platform written in Java that can be imported as a Python or R library	https://h2o.ai
Keras	High-level neural-network API written in Python	https://keras.io
Mlpack	Scalable machine-learning library written in C++	https://mlpack.org
Scikit-learn	Machine-learning and data-mining member of the scikit family of toolboxes built around the SciPy Python library	http://scikit-learn.org
Weka	Collection of machine-learning algorithms and tasks written in Java	https://cs.waikato.ac.nz/ml/weka
Machine-learning tools for molecules and materials		
Amp	Package to facilitate machine learning for atomistic calculations	https://bitbucket.org/andrewpeterson/amp
ANI	Neural-network potentials for organic molecules with Python interface	https://github.com/isayev/ASE_ANI
COMBO	Python library with emphasis on scalability and efficiency	https://github.com/tsudalab/combo
DeepChem	Python library for deep learning of chemical systems	https://deepchem.io
GAP	Gaussian approximation potentials	http://libatoms.org/Home/Software
MatMiner	Python library for assisting machine learning in materials science	https://hackingmaterials.github.io/matminer
NOMAD	Collection of tools to explore correlations in materials datasets	https://analytics-toolkit.nomad-coe.eu
PROPhet	Code to integrate machine-learning techniques with quantum-chemistry approaches	https://github.com/biklooost/PROPhet
TensorMol	Neural-network chemistry package	https://github.com/jparkhill/TensorMol

per step can range from around 80 to several thousand[30]; for comparison, there are only tens of potential moves at each game position in chess[31]. In chemical synthesis, human experts are required to specify conditional and contextual rules, which exclude large sets of potential reagents at a given step, limiting the number of choices available to the algorithm. The contextual rules (typically many thousands of them) are of utmost importance if a machine relying on a traditional algorithm is to compete with an expert. Recent breakthroughs in the Chematica program have shown that computers can be more efficient than humans in these tasks[32].

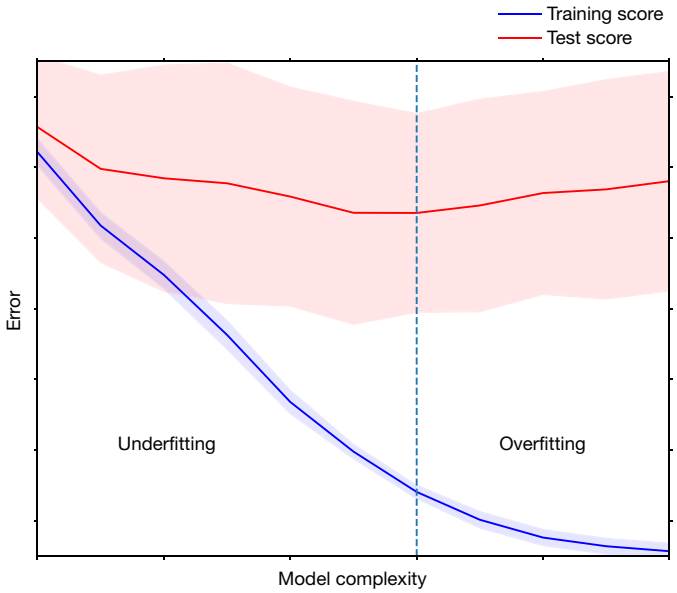


Fig. 2 | Errors that arise in machine-learning approaches. Errors can arise during both the training of a new model (blue line) and the application of a built model (red line). A simple model may suffer from high bias (underfitting), whereas a complex model may suffer from high variance (overfitting), which leads to a bias–variance trade-off. In the underfitting region the model performance can improve with further parameterization, whereas in the overfitting region the model performance will decrease. The optimal point for a model is just before the performance on the testing set starts to deteriorate with increased parameterization, which is indicated by the dashed vertical line. The model shown here is built on an example from https://kaggle.com, available at https://keeeto.github.io/blog/bias_variance. The shaded areas show the standard deviations of the fits for model training (blue) and testing (red).

The combination of extremely complex systems and huge numbers of potential solutions, which arise from competing objective functions (such as cost, purity, time and toxicity), make synthetic chemistry ill-suited to the application of traditional algorithmic approaches. However, because of this complexity, synthesis is one area of research that can benefit most from the application of artificial intelligence.

Deep-learning approaches, which typically rely on many-layered artificial neural networks or a combination of artificial neural networks and other learning techniques such as Boltzmann machines, are showing particular promise for predicting chemical-synthesis routes by combining rules-based expert systems with neural networks that rank the candidate synthetic pathways[33] or the likelihood of a predicted product by applying the rules[34]. One artificial neural network that learned from examples taken from the chemical literature was able to achieve a level of sophistication such that trained chemists could not distinguish between computer- and human-expert-designed routes[30]. However, a severe drawback of rules-based systems is that they have difficulty operating outside their knowledge base.

Alternatives to rules-based synthesis prediction have also been proposed, for example, so-called sequence-to-sequence approaches, which are based on the relationships between organic chemistry and linguistics. By casting molecules as text strings, these relationships have been applied in several chemical-design studies[35,36]. In sequence-to-sequence approaches, a model is fed an input of products and then outputs reactants as a SMILES string[37]. A similar approach has also been applied to retrosynthesis[38]. Future developments in areas such as one-shot learning (as recently applied to drug discovery)[39] could lead to wider application of non-rules-based methods in fields such as natural product synthesis, for which training data are scarce.

Beyond the synthesis of a target molecule, machine-learning models can be applied to assess the likelihood that a product will crystallize. By applying feature-selection techniques, a two-parameter model capable of predicting the propensity of a given molecule to crystallize

© 2018 Springer Nature Limited. All rights reserved.

with an accuracy of around 80% has been demonstrated[40]. Crucially, this model had access to a training set of more than 20,000 crystalline and non-crystalline compounds. The availability of such open-access databases is pivotal for the further development of similar predictive models[41]. In another study, a model was trained to predict the reaction conditions for new organically templated inorganic-product formation with a success rate of 89%[42].

A less explored avenue of machine learning is how to best sample the set of possible experimental set-ups. Active learning predicts the optimal future experiments that are required to better understand a given problem. It was recently applied to help to understand the conditions for the synthesis and crystallization of complex polyoxometalate clusters[43]. Starting from initial data on failed and successful experiments, the machine-learning approach directed future experiments and was shown to be capable of covering six times as much crystallization space as a human researcher in the same number of experiments.

Computational assistance for the planning and direction of chemical synthesis has come a long way since the early days of hand-coded expert systems. Much of this progress has been achieved in the past five years. Incorporation of artificial-intelligence-based chemical planners, with advances in robotic synthesis[43], promises a rich new frontier in the production of novel compounds.

Assisting multi-dimensional characterization

The structure of molecules and materials is typically deduced by a combination of experimental methods, such as X-ray and neutron diffraction, magnetic and spin resonance, and vibrational spectroscopy. Each approach has a certain sensitivity and length scale, and information from each method is complementary. Unfortunately, it is rare that data are fully assimilated into a coherent description of atomic structure. Analyses of individual streams often result in conflicting descriptions of the same compound[44]. A solution could be to incorporate real-time data into the modelling, with results then returned to the experiment, forming a feedback loop[45]. Machine learning represents a unifying framework that could enable the synergy of synthesis, imaging, theory and simulations.

The power of machine-learning methods for enhancing the link between modelling and experiment has been demonstrated in the field of surface science. Complex surface reconstructions have been characterized by combining ab initio simulations with multi-stage pattern-recognition systems that use convolutional neural networks[46]. Machine-learning methods have also recently shown promise in areas such as microstructural characterization[47] and the identification of interesting regions in large, complex, neutron-scattering, volumetric (three-dimensional) datasets[48]. Another example of machine learning opening new avenues in an area of complicated characterization is phase transitions of highly correlated systems; neural networks have been trained to encode topological phases of matter and thus identify transitions between them[49].

Enhancing theoretical chemistry

Modelling is now commonly considered as being equally important as synthesis and characterization for successful programmes of research. Using atomistic simulations, the properties of a molecule or material can, in principle, be calculated for any chemical composition and atomic structure. In practice, the computations grow rapidly in complexity as the size of the system increases, so considerable effort is devoted to finding short-cuts and approximations that enable the properties of the material to be calculated to an acceptable degree of fidelity, without the need for unreasonable amounts of computer time.

Approaches based on DFT have been successful in predicting the properties of many classes of compounds, offering generally high accuracy at reasonable cost. However, DFT and related electronic-structure techniques are limited by the exchange-correlation functional that describes non-classical interactions between electrons. There are notable limitations of the current approximations for weak chemical interactions (such as in layered materials), for highly correlated (*d* and *f* electron) systems and for the latest generation of quantum materials (such as iron pnictide superconductors), which often require a more sophisticated many-body Hamiltonian. Drawing from the growing number of structure–property databases (Table 3), accurate universal density functionals can be learned from data[50,51]. Early examples include the Bayesian error-estimation functional[52] and combinatorially optimized DFT functionals[53]. Going beyond the standard approach to DFT, the need to solve the Kohn–Sham equations can be by-passed by learning density-to-energy and density-to-potential maps directly from training systems[54].

Equally challenging is the description of chemical processes across length scales and timescales, such as the corrosion of metals in the presence of oxygen and water. A realistic description of chemical interactions (bond forming and breaking) including solvents, interfaces and disorder is still limited by the computational cost of available quantum-mechanical approaches. The task of developing transferrable analytic force fields is a well-defined problem for machine learning[55,56]. It has been demonstrated that, in simple materials, approximate potential-energy surfaces learned from quantum-mechanical data can save orders of magnitude in processing cost[57,58]. Although the combination of methods with varying levels of approximation is promising, much work is needed in the quantification and minimization of error propagation across methods. In this context, initiatives for error estimation such as the DAKOTA package (https://dakota.sandia.gov) are critically important.

Targeting discovery of new compounds

We have considered how machine learning can be used to enhance and integrate synthesis, characterization and modelling. However, machine learning can also reveal new ways of discovering compounds. Models that relate system descriptors to desirable properties are already used to reveal previously unknown structure–property relationships[59,60]. So far, the fields of molecular (primarily pharmaceutical and medicinal) and materials chemistry have experienced different degrees of uptake of machine-learning approaches to the design of new compounds, in part owing to the challenges of representing the crystal structure and morphology of extended solids.

Crystalline solids. The application of machine learning to the discovery of functional materials is an emerging field. An early report in 1998 applied machine learning to the prediction of magnetic and optoelectronic materials[61], but the number of studies has risen substantially only since 2010[62–64]. The complexity of games like Go is reminiscent of certain problems in materials science[65,66], such as the description of on-lattice interactions that govern chemical disorder, magnetism and ferroelectricity. Even for representations of small unit cells, the number of configurations of a disordered crystal can quickly exceed the limitations of conventional approaches. An inverse-design procedure illustrated how such a combinatorial space for an alloy could be harnessed to realize specific electronic structure features[67]. Similar inverse-design approaches have also been applied in molecular chemistry to tailor ground- and excited-state properties[68].

Predicting the likelihood of a composition to adopt a given crystal structure is a good example of a supervised classification problem in machine learning. Two recent examples involve predicting how likely a given composition is to adopt the so-called Heusler and half-Heusler crystal structures. The first predicts the likelihood that a given composition will adopt the Heusler structure and is trained on experimental data[69]. This approach was applied to screen hypothetical compositions and successfully identified 12 new gallide compounds, which were subsequently verified experimentally. In the second, a random forest model was trained on experimental data to learn the probability that a given ABC stoichiometry would adopt the half-Heusler structure[70].

As an alternative to learning from experimental data, calculated properties can be used as a training set for machine learning. Assessing the degree of similarity between electronic band structures has been shown to yield improved photocathodes for dye-sensitized solar cells[71]. A machine-learning model, trained to reproduce energies for the

© 2018 Springer Nature Limited. All rights reserved.

Table 3 | Publicly accessible structure and property databases for molecules and solids

Name	Description	URL
Computed structures and properties		
AFLOWLIB	Structure and property repository from high-throughput ab initio calculations of inorganic materials	http://aflowlib.org
Computational Materials Repository	Infrastructure to enable collection, storage, retrieval and analysis of data from electronic-structure codes	https://cmr.fysik.dtu.dk
GDB	Databases of hypothetical small organic molecules	http://gdb.unibe.ch/downloads
Harvard Clean Energy Project	Computed properties of candidate organic solar absorber materials	https://cepdb.molecularspace.org
Materials Project	Computed properties of known and hypothetical materials carried out using a standard calculation scheme	https://materialsproject.org
NOMAD	Input and output files from calculations using a wide variety of electronic-structure codes	https://nomad-repository.eu
Open Quantum Materials Database	Computed properties of mostly hypothetical structures carried out using a standard calculation scheme	http://oqmd.org
NREL Materials Database	Computed properties of materials for renewable-energy applications	https://materials.nrel.gov
TEDesignLab	Experimental and computed properties to aid the design of new thermo-electric materials	http://tedesignlab.org
ZINC	Commercially available organic molecules in 2D and 3D formats	https://zinc15.docking.org
Experimental structures and properties		
ChEMBL	Bioactive molecules with drug-like properties	https://www.ebi.ac.uk/chembl
ChemSpider	Royal Society of Chemistry's structure database, featuring calculated and experimental properties from a range of sources	https://chemspider.com
Citrination	Computed and experimental properties of materials	https://citrination.com
Crystallography Open Database	Structures of organic, inorganic, metal–organic compounds and minerals	http://crystallography.net
CSD	Repository for small-molecule organic and metal–organic crystal structures	https://www.ccdc.cam.ac.uk
ICSD	Inorganic Crystal Structure Database	https://icsd.fiz-karlsruhe.de
MatNavi	Multiple databases targeting properties such as superconductivity and thermal conductance	http://mits.nims.go.jp
MatWeb	Datasheets for various engineering materials, including thermoplastics, semi-conductors and fibres	http://matweb.com
NIST Chemistry WebBook	High-accuracy gas-phase thermochemistry and spectroscopic data	https://webbook.nist.gov/chemistry
NIST Materials Data Repository	Repository to upload materials data associated with specific publications	https://materialsdata.nist.gov
PubChem	Biological activities of small molecules	https://pubchem.ncbi.nlm.nih.gov

elpasolite crystal structure (ABC_2D_6), was applied to screen all two million possible combinations of elements that satisfy the formula, revealing chemical trends and identifying 128 new materials[72]. Such models are expected to become a central feature in the next generation of high-throughput virtual screening procedures.

The majority of crystal-solid machine-learning studies so far have concentrated on a particular type of crystal structure. This is because of the difficulty of representing crystalline solids in a format that can be fed easily to a statistical learning procedure. By concentrating on a single structure type, the representation is inherently built into the model. Developing flexible, transferrable representations is one of the important areas of research in machine learning for crystalline solids (see subsection 'Data representation'). As we will see below, the use of machine learning in molecular chemistry is more advanced than in the solid state, to a large extent owing to the greater ease with which molecules can be described in a manner amenable to algorithmic interpretation.

Molecular science. The quantitative structure–activity relationship is now a firmly established tool for drug discovery and molecular design. With the development of massive databases of assayed and virtual molecules[73,74], methods for rapid, reliable, virtual screening of these molecules for pharmacological (or other) activity are required to unlock the potential of such molecules. Models based on quantitative structure–activity relationships can be described as the application of statistical methods to the problem of finding empirical relationships of the type $P_i = k'(D_1, D_2, \ldots, D_n)$, where P_i is the property of interest, k' is a (typically linear) mathematical transformation and D_i are calculated or measured structural properties[75]. Machine learning has a long history in the development of quantitative structure–activity relationships, stretching back over half a century[76].

Molecular science is benefitting from cutting-edge algorithmic developments in machine learning such as generative adversarial networks[77] and reinforcement learning for the computational design of biologically active compounds. In a generative adversarial network, two models are trained simultaneously: a generative model (or generator) captures the distribution of data while a discriminative model (or discriminator) estimates the probability that a sample came from the training set rather than the generator. The training procedure for the generator is to maximize the probability of the discriminator making an error (Fig. 3). Models based on objective-reinforced generative adversarial networks[78] are capable of generating new organic molecules from scratch. Such models can be trained to produce diverse molecules that contain specific chemical features and physical responses, through a reward mechanism that resembles classical conditioning in psychology. Using reinforcement learning, newly generated chemical structures can be biased towards those with the desired physical and biological properties (de novo design).

Reclaiming the literature

A final area for which we consider the recent progress of machine learning (across all disciplines) is tapping into the vast amount of knowledge that already exists. Although the scientific literature provides a wealth of information to researchers, it is increasingly difficult to navigate owing to the proliferation of journals, articles and databases. Text mining has become a popular approach to identifying and extracting

© 2018 Springer Nature Limited. All rights reserved.

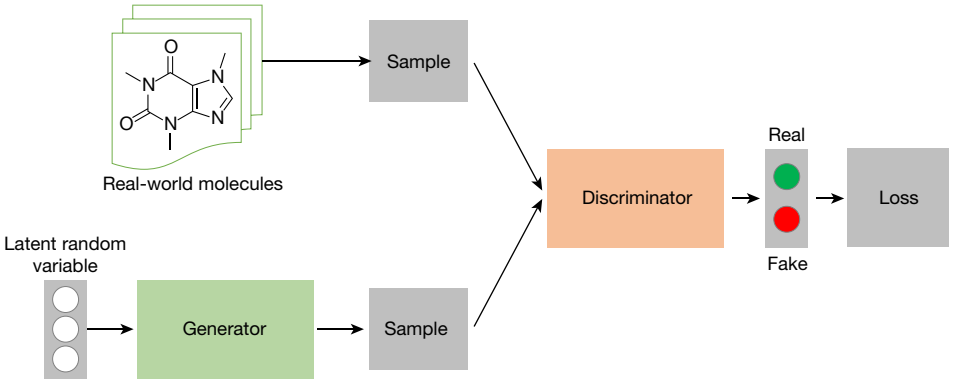


Fig. 3 | The generative adversarial network (GAN) approach to molecular discovery. Two models—a generator and a discriminator—play a continuous 'game'. The generator learns to map from a latent random variable (noise) to a particular distribution of molecules. The discriminator learns to get better and better at distinguishing fake data from real data. The two artificial neural networks are optimizing a different and opposing objective function, or loss function, in a zero-sum game. The general mathematical formulation for the GAN approach to unsupervised machine learning is outlined in ref. [77].

information from unstructured text sources. This approach can be used to extract facts and relationships in a structured form to create specialized databases, to transfer knowledge between domains and, more generally, to support research decision-making[79]. Text mining is applied to answer many different research questions, including in the discovery of drug–protein target associations, the analysis of high-throughput experiments and the development of systematic materials databases[80]. Owing to the heterogeneous nature of written resources, the automated extraction of relevant information is far from trivial. To address this, text mining has evolved into a sophisticated and specialized field where text-processing and machine-learning techniques are combined.

In cases where supplementary data are provided with a publication, it is made available in various formats and databases, often without validated or standardized metadata. The issue of data and metadata interoperability is key. Some examples of forward-looking initiatives that are pushing accessible, reusable data in scientific research are The Molecular Sciences Software Institute (http://molssi.org) and the Open Science Monitor (https://ec.europa.eu/research/openscience).

Frontiers in machine learning

There are many opportunities for further breakthroughs in machine learning to provide even greater advances in the automated design and discovery of molecules and materials. Here we highlight some frontiers in the field.

More knowledge from smaller datasets

Machine-learning approaches typically require large amounts of data for learning to be effective. Although this is rarely an issue in fields such as image recognition, in which millions of input datasets are available, in chemistry or materials science we are often limited to hundreds or thousands, if not fewer, high-quality data points. Researchers need to become better at making the data associated with our publications accessible in a computer-readable form. Another promising solution to the problem of limited datasets is meta-learning, whereby knowledge is learned within and across problems[81]. New developments such as neural Turing machines[82] and imitation learning[83] are enabling the realization of this process. A Bayesian framework has recently been reported to achieve human-level performance on one-shot learning problems with limited data[84], which has consequences for molecular and materials science where data are sparse and generally expensive and slow to obtain.

Efficient chemical representations

The standard description of chemical reactions, in terms of composition, structure and properties, has been optimized for human learning. Most machine-learning approaches for chemical reactions or properties use molecular or atomic descriptors to build models, the success of which is determined by the validity and relevance of these descriptors. A good descriptor must be simpler to obtain than the target property and of as low dimensionality as possible[85]. In the context of materials, useful descriptors[86] and approaches for adapting simple existing heuristics for machine learning have been outlined[87]; however, much work remains to develop powerful new descriptions. In the field of molecular reactions, advances such as the use of neural networks to create fingerprints for molecules in reactions are leading to improvements in synthesis prediction[88]. As has been demonstrated by the successful adoption of the concept of molecular fragments[23], the field of crystalline-materials design can learn much from advances in molecular nomenclature and representation.

Quantum learning

Whereas classical computing processes bits that are either 1 or 0, quantum computers use the quantum superposition of states to process qubits that are both 1 and 0 at the same time[89]. This parallelization leads to an exponential speedup in computational efficiency as the number of (qu)bits used increases[90]. Computational chemistry is a strong candidate to benefit from quantum computing because solving Schrödinger's equation on a quantum computer that consists of interacting electrons has a natural fit[91]. One of the challenges for quantum computing is knowing how to detect and correct errors that may occur in the data. Despite substantial efforts in industry and academia, no error-corrected qubits have been built so far. Quantum machine learning explores the application of machine-learning approaches to quantum problems, and vice versa—the application of quantum computing to machine-learning problems. The possibility of exponential speedups in optimization problems means that quantum machine learning has enormous potential. In problems such as optimizing synthetic routes[92] or improving a given metric (for example, optical absorption for solar energy materials) where multiple acceptable solutions exist, loss of qubit fidelity is less serious than when certainty is required. The physical sciences could prove a particularly rich field for quantum-learning applications[93,94].

Establishing new principles

Automatic discovery of scientific laws and principles[95,96] by inspection of the weights of trained machine-learning systems is a potentially transformational development in science. Although models developed from machine learning are predictive, they are not necessarily (or even usually) interpretable; there are several reasons for this. First, the way in which a machine-learning model represents knowledge rarely maps directly to forms that scientists are familiar with. Given suitable data, an artificial neural network might discover the ideal gas law ($pV = nRT$), but the translation of connection weights into a formula, typically through statistical learning, is non-trivial, even for a simple law such as this. The second reason is more subtle: the laws that underlie the

© 2018 Springer Nature Limited. All rights reserved.

behaviour of a compound might depend on knowledge that scientists do not yet possess, such as a many-body interaction giving rise to a new type of superconductivity. If an advanced machine-learning system was able to learn key aspects of quantum mechanics, it is hard to envisage how its connection weights could be turned into a comprehensible theory if the scientist lacked understanding of a fundamental component of it. Finally, there may be scientific laws that are so complex that, were they to be discovered by a machine-learning system, they would be too challenging for even a knowledgeable scientist to understand. A machine-learning system that could discern and use such laws would truly be a computational black box.

As scientists embrace the inclusion of machine learning with statistically driven design in their research programmes, the number of reported applications is growing at an extraordinary rate. This new generation of computational science, supported by a platform of open-source tools and data sharing, has the potential to revolutionize molecular and materials discovery.

Received: 20 October 2017; Accepted: 9 May 2018;
Published online 25 July 2018.

1. Dirac, P. A. M. Quantum mechanics of many-electron systems. *Proc. R. Soc. Lond. A* **123**, 714–733 (1929).
2. Pople, J. A. Quantum chemical models (Nobel lecture). *Angew. Chem. Int. Ed.* **38**, 1894–1902 (1999).
3. Boyd, D. B. Quantum chemistry program exchange, facilitator of theoretical and computational chemistry in pre-internet history. *ACS Symp. Ser.* **1122**, 221–273 (2013).
4. Arita, M., Bowler, D. R. & Miyazaki, T. Stable and efficient linear scaling first-principles molecular dynamics for 10000+ atoms. *J. Chem. Theory Comput.* **10**, 5419–5425 (2014).
5. Wilkinson, K. A., Hine, N. D. M. & Skylaris, C.-K. Hybrid mpi-openmp parallelism in the Onetep linear-scaling electronic structure code: application to the delamination of cellulose nanofibrils. *J. Chem. Theory Comput.* **10**, 4782–4794 (2014).
6. Havu, V., Blum, V., Havu, P. & Scheffler, M. Efficient *O*(*N*) integration for all-electron electronic structure calculation using numeric basis functions. *J. Comput. Phys.* **228**, 8367–8379 (2009).
7. Catlow, C. R. A., Sokol, A. A. & Walsh, A. *Computational Approaches to Energy Materials* (Wiley-Blackwell, New York, 2013).
8. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136**, B864–B871 (1964).
9. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133–A1138 (1965).
10. Lejaeghere, K. et al. Reproducibility in density functional theory calculations of solids. *Science* **351**, aad3000 (2016).
11. Hachmann, J. et al. The Harvard clean energy project: large-scale computational screening and design of organic photovoltaics on the world community grid. *J. Phys. Chem. Lett.* **2**, 2241–2251 (2011).
12. Jain, A. et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
13. Calderon, C. E. et al. The AFLOW standard for high-throughput materials science calculations. *Comput. Mater. Sci.* **108**, 233–238 (2015).
14. Agrawal, A. & Choudhary, A. Perspective: Materials informatics and big data: realization of the 'fourth paradigm' of science in materials science. *APL Mater.* **4**, 053208 (2016).
15. Schwab, K. The fourth industrial revolution. *Foreign Affairs* https://www.foreignaffairs.com/articles/2015-12-12/fourth-industrial-revolution (2015).
16. Fourches, D., Muratov, E. & Tropsha, A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model.* **50**, 1189–1204 (2010).
17. Kireeva, N. et al. Generative topographic mapping (GTM): universal tool for data visualization, structure-activity modeling and dataset comparison. *Mol. Inform.* **31**, 301–312 (2012).
18. Faber, F. A. et al. Prediction errors of molecular machine learning models lower than hybrid DFT error. *J. Chem. Theory Comput.* **13**, 5255–5264 (2017).
19. Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
20. Bonchev, D. & Rouvray, D. H. *Chemical Graph Theory: Introduction and Fundamentals* (Abacus Press, New York, 1991).
21. Schütt, K. T. et al. How to represent crystal structures for machine learning: towards fast prediction of electronic properties. *Phys. Rev. B* **89**, 205118 (2014).
A radial-distribution-function description of periodic solids is adapted for machine-learning models and applied to predict the electronic density of states for a range of materials.
22. Ward, L. et al. Including crystal structure attributes in machine learning models of formation energies via Voronoi tessellations. *Phys. Rev. B* **96**, 024104 (2017).
23. Isayev, O. et al. Universal fragment descriptors for predicting electronic properties of inorganic crystals. *Nat. Commun.* **8**, 15679 (2017).
24. Hand, D. J. & Yu, K. Idiot's Bayes—not so stupid after all? *Int. Stat. Rev.* **69**, 385–398 (2001).
25. Shakhnarovich, G., Darrell, T. & Indyk, P. *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice* (MIT Press, Boston, 2005).
26. Rokach, L. & Maimon, O. in *Data Mining and Knowledge Discovery Handbook* (eds Maimon, O. & Rokach, L.) 149–174 (Springer, New York, 2010).
27. Shawe-Taylor, J. & Cristianini, N. *Kernel Methods for Pattern Analysis* (Cambridge Univ. Press, Cambridge, 2004).
28. Schmidhuber, J. Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117 (2015).
29. Corey, E. J. & Wipke, W. T. Computer-assisted design of complex organic synthesis. *Science* **166**, 178–192 (1969).
30. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
A computer-driven retrosynthesis tool was trained on most published reactions in organic chemistry.
31. Szymkuć, S. et al. Computer-assisted synthetic planning: the end of the beginning. *Angew. Chem. Int. Ed.* **55**, 5904–5937 (2016).
32. Klucznik, T. et al. Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. *Chem* **4**, 522–532 (2018).
33. Segler, M. H. S. & Waller, M. P. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chem. Eur. J.* **23**, 5966–5971 (2017).
34. Cole, J. C. et al. Generation of crystal structures using known crystal structures as analogues. *Acta Crystallogr. B* **72**, 530–541 (2016).
35. Gómez-Bombarelli, R. et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nat. Mater.* **15**, 1120–1127 (2016).
This study uses machine learning to guide all stages of a materials discovery workflow from quantum-chemical calculations to materials synthesis.
36. Jastrzębski, S., Leśniak, D. & Czarnecki, W. M. Learning to SMILE(S). Preprint at https://arxiv.org/abs/1602.06289 (2016).
37. Nam, J. & Kim, J. Linking the neural machine translation and the prediction of organic chemistry reactions. Preprint at https://arxiv.org/abs/1612.09529 (2016).
38. Liu, B. et al. Retrosynthetic reaction prediction using neural sequence-to-sequence models. *ACS Cent. Sci.* **3**, 1103–1113 (2017).
39. Altae-Tran, H., Ramsundar, B., Pappu, A. S. & Pande, V. Low data drug discovery with one-shot learning. *ACS Cent. Sci.* **3**, 283–293 (2017).
40. Wicker, J. G. P. & Cooper, R. I. Will it crystallise? Predicting crystallinity of molecular materials. *CrystEngComm* **17**, 1927–1934 (2015).
This paper presents a crystal engineering application of machine learning to assess the probability of a given molecule forming a high-quality crystal.
41. Pillong, M. et al. A publicly available crystallisation data set and its application in machine learning. *CrystEngComm* **19**, 3737–3745 (2017).
42. Raccuglia, P. et al. Machine-learning-assisted materials discovery using failed experiments. *Nature* **533**, 73–76 (2016).
The study trains a machine-learning model to predict the success of a chemical reaction, incorporating the results of unsuccessful attempts as well as known (successful) reactions.
43. Dragone, V., Sans, V., Henson, A. B., Granda, J. M. & Cronin, L. An autonomous organic reaction search engine for chemical reactivity. *Nat. Commun.* **8**, 15733 (2017).
44. Billinge, S. J. L. & Levin, I. The problem with determining atomic structure at the nanoscale. *Science* **316**, 561–565 (2007).
45. Kalinin, S. V., Sumpter, B. G. & Archibald, R. K. Big–deep–smart data in imaging for guiding materials design. *Nat. Mater.* **14**, 973–980 (2015).
46. Ziatdinov, M., Maksov, A. & Kalinin, S. V. Learning surface molecular structures via machine vision. *npj Comput. Mater.* **3**, 31 (2017).
47. de Albuquerque, V. H. C., Cortez, P. C., de Alexandria, A. R. & Tavares, J. M. R. S. A new solution for automatic microstructures analysis from images based on a backpropagation artificial neural network. *Nondestruct. Test. Eval.* **23**, 273–283 (2008).
48. Hui, Y. & Liu, Y. Volumetric data exploration with machine learning-aided visualization in neutron science. Preprint at https://arxiv.org/abs/1710.05994 (2017).
49. Carrasquilla, J. & Melko, R. G. Machine learning phases of matter. *Nat. Phys.* **13**, 431–434 (2017).
50. Christensen, R., Hansen, H. A. & Vegge, T. Identifying systematic DFT errors in catalytic reactions. *Catal. Sci. Technol.* **5**, 4946–4949 (2015).
51. Snyder, J. C., Rupp, M., Hansen, K., Müller, K.-R. & Burke, K. Finding density functionals with machine learning. *Phys. Rev. Lett.* **108**, 253002 (2012).
52. Wellendorff, J. et al. Density functionals for surface science: exchange-correlation model development with Bayesian error estimation. *Phys. Rev. B* **85**, 235149 (2012).
53. Mardirossian, N. & Head-Gordon, M. ωB97M-V a combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation. *J. Chem. Phys.* **144**, 214110 (2016).
54. Brockherde, F. et al. Bypassing the Kohn-Sham equations with machine learning. *Nat. Commun.* **8**, 872 (2017).
This study transcends the standard approach to DFT by providing a direct mapping from density to energy, paving the way for higher-accuracy approaches.
55. Behler, J. First principles neural network potentials for reactive simulations of large molecular and condensed systems. *Angew. Chem. Int. Ed.* **56**, 12828–12840 (2017).

© 2018 Springer Nature Limited. All rights reserved.

56. Smith, J. S., Isayev, O. & Roitberg, A. E. Ani-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **8**, 3192–3203 (2017).
57. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: the accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
In this study, machine learning is used to fit interatomic potentials that reproduce the total energy and energy derivatives from quantum-mechanical calculations and enable accurate low-cost simulations.
58. Handley, C. M. & Popelier, P. L. A. Potential energy surfaces fitted by artificial neural networks. *J. Phys. Chem. A* **114**, 3371–3383 (2010).
59. Pulido, A. et al. Functional materials discovery using energy–structure–function maps. *Nature* **543**, 657–664 (2017).
60. Hill, J. et al. Materials science with large-scale data and informatics: unlocking new opportunities. *MRS Bull.* **41**, 399–409 (2016).
61. Kiselyova, N. N., Gladun, V. P. & Vashchenko, N. D. Computational materials design using artificial intelligence methods. *J. Alloys Compd.* **279**, 8–13 (1998).
62. Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *JOM* **65**, 1501–1509 (2013).
63. Pilania, G., Wang, C., Jiang, X., Rajasekaran, S. & Ramprasad, R. Accelerating materials property predictions using machine learning. *Sci. Rep.* **3**, 2810 (2013).
64. Hautier, G., Fischer, C. C., Jain, A., Mueller, T. & Ceder, G. Finding nature's missing ternary oxide compounds using machine learning and density functional theory. *Chem. Mater.* **22**, 3762–3767 (2010).
In an early example of harnessing materials databases, information on known compounds is used to construct a machine-learning model to predict the viability of previously unreported chemistries.
65. Walsh, A. The quest for new functionality. *Nat. Chem.* **7**, 274–275 (2015).
66. Davies, D. W. et al. Computational screening of all stoichiometric inorganic materials. *Chem* **1**, 617–627 (2016).
67. Franceschetti, A. & Zunger, A. The inverse band-structure problem of finding an atomic configuration with given electronic properties. *Nature* **402**, 60–63 (1999).
68. Kuhn, C. & Beratan, D. N. Inverse strategies for molecular design. *J. Phys. Chem.* **100**, 10595–10599 (1996).
69. Oliynyk, A. O. et al. High-throughput machine-learning-driven synthesis of full-Heusler compounds. *Chem. Mater.* **28**, 7324–7331 (2016).
70. Legrain, F., Carrete, J., van Roekeghem, A., Madsen, G. K. H. & Mingo, N. Materials screening for the discovery of new half-heuslers: machine learning versus ab initio methods. *J. Phys. Chem. B* **122**, 625–632 (2018).
71. Moot, T. et al. Material informatics driven design and experimental validation of lead titanate as an aqueous solar photocathode. *Mater. Discov.* **6**, 9–16 (2016).
72. Faber, F. A., Lindmaa, A., Von Lilienfeld, O. A. & Armiento, R. Machine learning energies of 2 million elpasolite (ABC_2D_6) crystals. *Phys. Rev. Lett.* **117**, 135502 (2016).
73. Oprea, T. I. & Tropsha, A. Target, chemical and bioactivity databases – integration is key. *Drug Discov. Today. Technol.* **3**, 357–365 (2006).
74. Sterling, T. & Irwin, J. J. ZINC 15 – ligand discovery for everyone. *J. Chem. Inf. Model.* **55**, 2324–2337 (2015).
75. Tropsha, A. Best practices for QSAR model development, validation, and exploitation. *Mol. Inform.* **29**, 476–488 (2010).
76. Hansch, C. & Fujita, T. p-σ-π analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* **86**, 1616–1626 (1964).
77. Goodfellow, I. J. et al. Generative adversarial networks. Preprint at https://arxiv.org/abs/1406.2661 (2014).
78. Guimaraes, G. L., Sanchez-Lengeling, B., Outeiral, C., Farias, P. L. C. & Aspuru-Guzik, A. Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models. Preprint at https://arxiv.org/abs/1705.10843 (2017).
79. Fleuren, W. W. M. & Alkema, W. Application of text mining in the biomedical domain. *Methods* **74**, 97–106 (2015).
80. Kim, E. et al. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem. Mater.* **29**, 9436–9444 (2017).
81. Jankowski, N., Duch, W. & Grąbczewski, K. (eds) *Meta-Learning in Computational Intelligence* (Springer, Berlin, 2011).
82. Graves, A., Wayne, G. & Danihelka, I. Neural Turing machines. Preprint at https://arxiv.org/abs/1410.5401 (2014).
83. Duan, Y. et al. One-shot imitation learning. Preprint at https://arxiv.org/abs/1703.07326 (2017).
84. Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).
85. Ghiringhelli, L. M., Vybiral, J., Levchenko, S. V., Draxl, C. & Scheffler, M. Big data of materials science: critical role of the descriptor. *Phys. Rev. Lett.* **114**, 105503 (2015).
86. Curtarolo, S. et al. The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191–201 (2013).
87. Seko, A., Togo, A. & Tanaka, I. in *Nanoinformatics* (ed. Tanaka, I.) 3–23 (Springer, Singapore, 2018).
88. Duvenaud, D. et al. Convolutional networks on graphs for learning molecular fingerprints. Preprint at https://arxiv.org/abs/1509.09292 (2015).
89. Steane, A. Quantum computing. *Rep. Prog. Phys.* **61**, 117 (1998).
90. Harrow, A. W., Hassidim, A. & Lloyd, S. Quantum algorithm for linear systems of equations. *Phys. Rev. Lett.* **103**, 150502 (2009).
91. Aspuru-Guzik, A., Dutoi, A. D., Love, P. J. & Head-Gordon, M. Simulated quantum computation of molecular energies. *Science* **309**, 1704–1707 (2005).
In an early application of quantum computing to molecular problems, a quantum algorithm that scales linearly with the number of basis functions is demonstrated for calculating properties of chemical interest.
92. Reiher, M., Wiebe, N., Svore, K. M., Wecker, D. & Troyer, M. Elucidating reaction mechanisms on quantum computers. *Proc. Natl Acad. Sci. USA* **114**, 7555–7560 (2017).
93. Dunjko, V., Taylor, J. M. & Briegel, H. J. Quantum-enhanced machine learning. *Phys. Rev. Lett.* **117**, 130501 (2016).
94. Biamonte, J. et al. Quantum machine learning. *Nature* **549**, 195–202 (2017).
95. Schmidt, M. & Lipson, H. Distilling free-form natural laws from experimental data. *Science* **324**, 81–85 (2009).
96. Rudy, S. H., Brunton, S. L., Proctor, J. L. & Kutz, J. N. Data-driven discovery of partial differential equations. *Sci. Adv.* **3**, e1602614 (2017).
97. Domingos, P. *The Master Algorithm* (Basic Books, New York, 2015).
98. Coudert, F.-X. Reproducible research in computational chemistry of materials. *Chem. Mater.* **29**, 2615–2617 (2017).
99. Tetko, I. V., Maran, U. & Tropsha, A. Public (Q)SAR services, integrated modeling environments, and model repositories on the web: state of the art and perspectives for future development. *Mol. Inform.* **36**, 1600082 (2017).

Acknowledgements This work was supported by the EPSRC (grant numbers EP/M009580/1, EP/K016288/1 and EP/L016354/1), the Royal Society and the Leverhulme Trust. O.I. acknowledges support from DOD-ONR (N00014-16-1-2311) and an Eshelman Institute for Innovation award.

Reviewer information *Nature* thanks F.-X. Coudert, M. Waller and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions All authors contributed equally to the design, writing and editing of the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to O.I. or A.W.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

ARTICLE

https://doi.org/10.1038/s41586-018-0347-0

Niobium tungsten oxides for high-rate lithium-ion energy storage

Kent J. Griffith[1], Kamila M. Wiaderek[2], Giannantonio Cibin[3], Lauren E. Marbella[1] & Clare P. Grey[1]*

The maximum power output and minimum charging time of a lithium-ion battery depend on both ionic and electronic transport. Ionic diffusion within the electrochemically active particles generally represents a fundamental limitation to the rate at which a battery can be charged and discharged. To compensate for the relatively slow solid-state ionic diffusion and to enable high power and rapid charging, the active particles are frequently reduced to nanometre dimensions, to the detriment of volumetric packing density, cost, stability and sustainability. As an alternative to nanoscaling, here we show that two complex niobium tungsten oxides—$Nb_{16}W_5O_{55}$ and $Nb_{18}W_{16}O_{93}$, which adopt crystallographic shear and bronze-like structures, respectively—can intercalate large quantities of lithium at high rates, even when the sizes of the niobium tungsten oxide particles are of the order of micrometres. Measurements of lithium-ion diffusion coefficients in both structures reveal room-temperature values that are several orders of magnitude higher than those in typical electrode materials such as $Li_4Ti_5O_{12}$ and $LiMn_2O_4$. Multielectron redox, buffered volume expansion, topologically frustrated niobium/tungsten polyhedral arrangements and rapid solid-state lithium transport lead to extremely high volumetric capacities and rate performance. Unconventional materials and mechanisms that enable lithiation of micrometre-sized particles in minutes have implications for high-power applications, fast-charging devices, all-solid-state energy storage systems, electrode design and material discovery.

New high-rate electrode materials that can store large quantities of charge in a few minutes, rather than hours, are required to increase power and decrease charging time in lithium-ion batteries. Such materials can thus help to alleviate technological challenges associated with the adoption of electric vehicles and grid-scale batteries and enable the development of new power-intensive devices. The most intuitive and commonly used approach to increase rate performance is to create nanometre-sized or porous (and often hierarchical) structures, which minimize Li^+ solid-state diffusion distances, enable more rapid Li^+ transport through the composite electrode and increase the surface area of electrode materials in contact with electrolyte. Carbonaceous hierarchical structures and carbon-coating are also frequently used to improve electronic conductivity, which is another prerequisite for the application of high current densities.

In practice, despite excellent lithium mobility, graphite cannot be used at high rates owing to particle fracture and the risk of Li dendrite formation, the latter leading to short circuits and potentially fires and explosions[1–3]. The dendrite issue inherently limits the use of low-voltage anodes in high-rate applications because electrode inhomogeneity, or any source of increased overpotential, can lead to Li plating potentials on the surface of the electrode[3]. $Li_4Ti_5O_{12}$, with an average voltage of 1.55 V (Supplementary Fig. 1), enables high-rate Li (de)intercalation without the risk of Li dendrites or substantial solid–electrolyte interphase formation, albeit with an undesirable but necessary decrease in full-cell voltage and thus energy density. In this well-established 'high'-voltage and high-rate anode, the capacity of 1-μm particles from solid-state synthesis reaches only 60–65 mA h g^{-1} at a rate[4] of 10C, where the C-rate is defined as the inverse of the number of hours required to reach a defined theoretical capacity; for example, 10C corresponds to a 6-min discharge or charge time (see Methods). By contrast, through two decades of research, present carbon-coated nanoparticles of $Li_4Ti_5O_{12}$ can reach at least 150 mA h g^{-1} at 10C[5,6], which corresponds to approximately 0.5 lithium ions per transition metal (Li^+/TM). However, using nanometre-sized and porous materials for electrochemical energy storage applications inherently results in a severe penalty in terms of volumetric energy density. Furthermore, these carefully designed porous and nano-architectures are time consuming and expensive to synthesize, characterize and manufacture. The relevant synthesis methods often result in relatively low yields or the generation of large quantities of chemical waste[7], while also being more susceptible to degradation during electrochemical cycling (from processes such as catalytic decomposition of the electrolyte[8], morphological changes that result in loss of nanostructuring[9] and higher first-cycle capacity loss[10]).

In this work, we break from the conventional strategy of nanoscaling and nanostructuring of electrode materials to overcome poor ionic diffusion and electronic properties (found in TiO_2 and $Li_4Ti_5O_{12}$ for example). We demonstrate that with the appropriate host lattice, none of the usual size, structure or porosity criteria are required to achieve a practical high-rate battery electrode. Instead, we use insight obtained from previous investigations of complex binary niobium oxides, such as the low-temperature polymorph T-Nb_2O_5[11], and of superionic conductors, such as lithium lanthanum titanate perovskite (LLTO)[12], to identify structural motifs that should exhibit favourable Li diffusion properties, and thus superior performance, allowing micrometre-sized particles to be used at extremely high rates. We show that when multi-redox 4*d* and 5*d* transition metals are used with the appropriate three-dimensional oxide structure, we can achieve extremely high volumetric energy densities and impressive rates. The bulk compounds studied are a series of complex 'block' or 'bronze-like' oxide structures (Fig. 1) largely comprising corner- and edge-sharing NbO_6 and WO_6 octahedra, and both oxides are prepared via gram-scale solid-state synthesis. Their unusual electrochemical performance is first illustrated by studying large (3–10 μm primary, 10–30 μm agglomerate) dense

[1]Department of Chemistry, University of Cambridge, Cambridge, UK. [2]X-ray Science Division, Advanced Photon Source, Argonne National Laboratory, Argonne, IL, USA. [3]Diamond Light Source, Harwell Science and Innovation Campus, Didcot, UK. *e-mail: cpg27@cam.ac.uk

© 2018 Springer Nature Limited. All rights reserved.

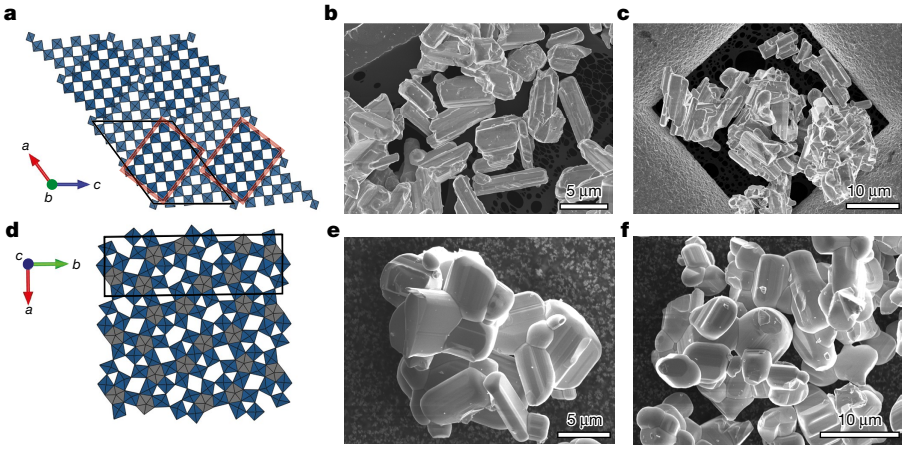


Fig. 1 | Crystal structure and particle morphology of $Nb_{16}W_5O_{55}$ and $Nb_{18}W_{16}O_{93}$. a–c, $Nb_{16}W_5O_{55}$ is built from blocks (red rectangles) of 4 × 5 (Nb,W)O_6 octahedra, with adjoining blocks forming crystallographic shear planes. (Nb,W)O_4 tetrahedra connect the corners of the blocks. **a**, A view down the *b* direction of the structure. **b**, **c** Electron images of the micrometre-sized particles. **d–f**, $Nb_{18}W_{16}O_{93}$ is a superstructure of the tetragonal tungsten bronze (blue) with pentagonal tunnels (grey) partially filled by –W–O– chains that form pentagonal bipyramids. **d**, A view down the *c* direction, depicting the various tunnels. **e**, **f**, Electron images. In **a** and **d**, the black boxes indicate the unit cells.

particles of the block structure $Nb_{16}W_5O_{55}$ (Fig. 1a–c). Even on a mass-normalized basis, the lithium storage performance of $Nb_{16}W_5O_{55}$ exceeds that of nanostructured versions of the heavily studied $Li_4Ti_5O_{12}$[5,13–15], TiO_2(B)[16–18] and T-Nb_2O_5[19–21] under similar loading conditions. Given the high density of the crystal structure and the high tap density of bulk $Nb_{16}W_5O_{55}$ compared with nanomaterials, this leads to exceptionally high volumetric performance. We further demonstrate the generality of this bulk phenomenon by exploring another new electrode material, bronze-like $Nb_{18}W_{16}O_{93}$ (Fig. 1d–f).

Complex oxides for energy storage

$Nb_{16}W_5O_{55}$ is a metastable member of the system Nb_2O_5–WO_3[22], with a monoclinic structure composed of subunits of corner-shared octahedra arranged into ReO_3-like blocks that are four octahedra wide by five octahedra long and infinite in the third dimension (Fig. 1a)[23]. The block subunits are connected by crystallographic shear planes along the edges and by tetrahedra at each corner, and the structure is expressed as $(4 \times 5)_1$; in the notation $(m \times n)_p$, *m* and *n* denote block length in units of octahedra and *p* relates to the connectivity of the blocks, which may also be joined in pairs ($p = 2$) or infinite ribbons ($p = \infty$). To the best of our knowledge, this is the first reported application of any kind for $Nb_{16}W_5O_{55}$ since its discovery in 1965[23,24]. $Nb_{18}W_{16}O_{93}$ is orthorhombic, a $1 \times 3 \times 1$ superstructure of the classic tetragonal tungsten bronze (Fig. 1d, Supplementary Fig. 2). The superstructure[25] results from partial filling of pentagonal tunnels by –M–O– chains to form pentagonal bipyramids, in addition to the distorted octahedra of the tetragonal tungsten bronzes. $Nb_{16}W_5O_{55}$ and $Nb_{18}W_{16}O_{93}$ were prepared via co-thermal oxidation of pellets of NbO_2 and WO_2. Details and alternative synthetic routes are described in Methods (Supplementary Figs. 3–7).

Reaction of $Nb_{16}W_5O_{55}$ with lithium (Fig. 2a) proceeds in three regions from 2.5 V to 1.0 V, with an average voltage of 1.57 V (Supplementary Fig. 1), comparable to the average voltage[26] of 1.55 V of $Li_4Ti_5O_{12}$. The three regions, more easily observed in the derivative plot (Fig. 2b), are characterized by their slope and are reminiscent of the three regions observed in other crystallographic shear structures[27], such as a second niobium oxide polymorph, H-Nb_2O_5[28], PNb_9O_{25}[29], $TiNb_2O_7$[30] and $Nb_{12}WO_{33}$[31]. When the kinetics were examined over a range of current densities from C/5 (34.3 mA g^{-1}) to 60C (10.3 A g^{-1}), $Nb_{16}W_5O_{55}$ showed unprecedented bulk rate performance (Fig. 2a, b, e) in standard electrode formulations (see Methods and Supplementary Fig. 8). At C/5, around 1.3 Li^+/TM can be reversibly intercalated for a gravimetric capacity of about 225 mA h g^{-1}. When the rate is increased by a factor of 25 to 5C, $Nb_{16}W_5O_{55}$ maintains a capacity of 1.0 Li^+/TM (171 mA h g^{-1}). At 20C, which corresponds to a three-minute discharge, it is still possible to exchange 0.86 Li^+/TM and access 148 mA h g^{-1}. Rate tests on $Nb_{16}W_5O_{55}$ were performed with a potentiostatic hold at the top of charge to ensure a reliable starting point for discharge. To test the performance under more demanding conditions, 1,000 cycles were measured with fixed galvanostatic discharge and charge conditions of 10C for 250 cycles followed by 20C for 750 cycles with no voltage hold (Fig. 2f). Under these conditions, 0.90 Li^+/TM (average 155 mA h g^{-1}) were reversibly intercalated at 10C with 95% capacity retention after 250 cycles on non-optimized or calendared electrodes. At 20C, the capacity was 0.75 Li^+/TM (average 128 mA h g^{-1}) and the capacity retention was again 95% over 750 cycles at 20C.

Excellent electrochemical energy storage was also discovered in another niobium tungsten oxide with distinct structural motifs: micrometre-scale particles of the bronze-like phase $Nb_{18}W_{16}O_{93}$ (Fig. 1d–f) showed enhanced rate performance and could be cycled at extremely high rates (Fig. 2c–f). The average voltage of $Nb_{18}W_{16}O_{93}$ is 1.67 V (Supplementary Fig. 1). In terms of gravimetric capacity, $Nb_{18}W_{16}O_{93}$ stores about 20 mA h g^{-1} less than $Nb_{16}W_5O_{55}$ at C/5 and 1C owing to the higher molar mass of the tungsten-rich bronze phase. However, at 20C, $Nb_{18}W_{16}O_{93}$ is still able to accommodate a full unit of Li^+/TM for a capacity of approximately 150 mA h g^{-1}. At 60C and 100C (14.9 A g^{-1}), the capacity is 105 and 70 mA h g^{-1}, respectively.

Other cycling conditions, such as long-term cycling at C/5, and the effect of current collectors, which cannot be ignored at high rates[32], were examined (Supplementary Fig. 9, 10). As a control, Li||Li symmetric cells were cycled at current densities corresponding to those applied for C/5–100C in Fig. 2 (Extended Data Fig. 1a). The overpotentials in the symmetric cell closely match those observed in the electrochemical cycling curves of Fig. 2a–d. This suggests that the extremely high rates in a bulk electrode approach the limits of Li metal plating/stripping or lithium-ion desolvation and transport in carbonate ester electrolytes at room temperature; that is, a considerable fraction of the ohmic loss during fast charging results from the Li metal and electrolyte rather than the complex oxide electrode materials.

Measuring diffusion coefficients in mixed ionic–electronic conductors

The extremely high mobility of Li^+ ions in these systems enabled the direct measurement of lithium diffusion with the pulsed-field-gradient nuclear magnetic resonance (PFG NMR) spectroscopy technique, which has previously only been applied to liquids or diamagnetic superionic solid electrolytes (Extended Data Table 1). To handle the short T_2 (spin–spin) relaxation times of ^{7}Li nuclei, which generally prevent these measurements in electrode materials because they are mixed ionic–electronic conductors, experiments were performed in the

© 2018 Springer Nature Limited. All rights reserved.

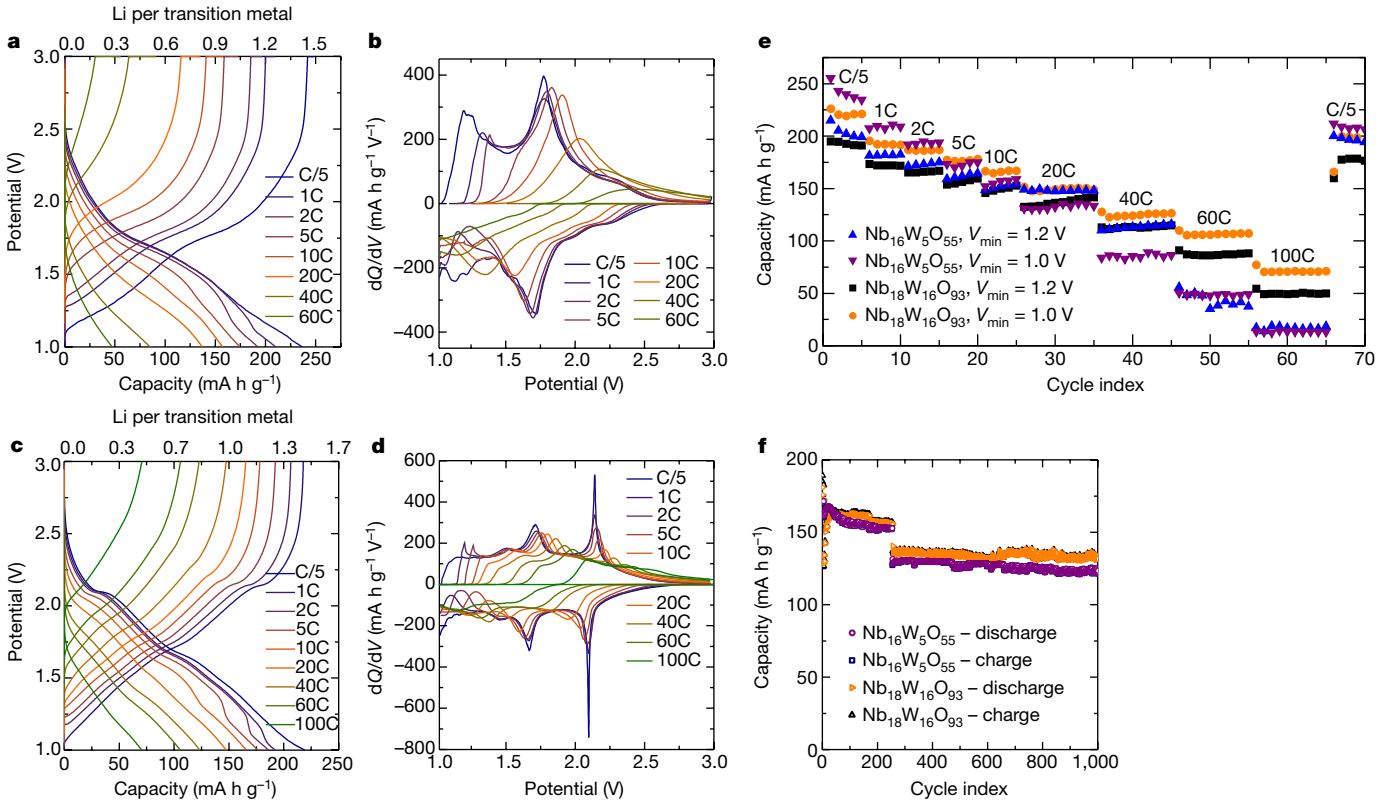


Fig. 2 | Electrochemistry of $Nb_{16}W_5O_{55}$ and $Nb_{18}W_{16}O_{93}$. **a**–**d**, Galvanostatic discharge and charge curves and dQ/dV plots (Q, capacity; V, voltage) of bulk $Nb_{16}W_5O_{55}$ (**a**, **b**) and $Nb_{18}W_{16}O_{93}$ (**c**, **d**) from C/5 up to 100C. **e**, Rate performance summary based on gravimetric capacity. **f**, High-rate cycling for 250 cycles at 10C, followed by 750 cycles at 20C. Dense electrodes of large particles with active mass loading of 2–3 mg cm^{-2} were tested at current densities corresponding to discharge times of several hours to tens of seconds. $Nb_{16}W_5O_{55}$ was charged with a 1-h constant-voltage step at the top of charge to ensure a comparable starting point on discharge; $Nb_{18}W_{16}O_{93}$ was cycled without this hold step for all measurements and stored over 100 mA h g^{-1} at 60C (that is, in less than 60 s). High-rate cycling for 1,000 cycles was performed on both oxides at 10C and 20C and at a constant current, without a potentiostatic step.

temperature range 333–453 K, at which the T_2 relaxation times were longer than at room temperature. Analysis of the data (Extended Data Fig. 2, Methods) for $Li_xNb_{16}W_5O_{55}$ ($x = 6.3$, 8.4) and $Li_xNb_{18}W_{16}O_{93}$ ($x = 3.4$, 6.8, 10.2) showed lithium transport above 10^{-13} m^2 s^{-1} at 333 K and 10^{-12} m^2 s^{-1} at 373 K. Assuming Arrhenius behaviour and the extremely low measured activation energies of 0.10–0.30 eV, the room-temperature lithium diffusion coefficients are estimated to be greater than 1×10^{-13} m^2 s^{-1} for all the materials (Table 1). The different samples exhibited similar diffusion coefficients and activation energies, consistent with the transport measurements made with the galvanostatic intermittent titration technique (GITT) (see Methods and Extended Data Fig. 1b, c) at low-to-moderate lithium contents. Lithium diffusion in both niobium tungsten oxide structures is markedly faster than that of $Li_{4+x}Ti_5O_{12}$ or Li_xTiO_2 (about 10^{-16}–10^{-15} m^2 s^{-1}) and is close to that of the best known lithium solid electrolytes (Extended Data Table 1). GITT and ^{7}Li PFG NMR spectroscopy results (Extended Data Figs. 1b and 2, respectively) indicate that this rapid motion is maintained to high lithium contents (≥1.5 Li$^+$/TM), and then the diffusion drops by about two orders of magnitude towards 2.0 Li$^+$/TM. These complementary techniques suggest that the inherent range of the niobium tungsten oxide electrode materials for high-rate multi-redox extends to approximately 1.5 Li$^+$/TM. The diffusion coefficients, of the order of 10^{-12}–10^{-13} m^2 s^{-1}, measured for these materials are consistent with the values required to achieve full lithiation of 10-μm particles on a 60C timescale (Supplementary Table 1).

Table 1 | Lithium diffusion coefficients, obtained from PFG NMR

Component	D_{Li} (m^2 s^{-1}) at 298 K	D_{Li} (m^2 s^{-1}) at 413 K	E_a (eV)
$Li_{6.3}Nb_{16}W_5O_{55}$-a (15%)	2.1×10^{-12}	2.7×10^{-11}	0.23 ± 0.04
$Li_{6.3}Nb_{16}W_5O_{55}$-b (85%)	1.7×10^{-13}	5.2×10^{-13}	0.10 ± 0.04
$Li_{8.4}Nb_{16}W_5O_{55}$	1.6×10^{-13}	6.5×10^{-13} (403 K)	0.13 ± 0.01
$Li_{3.4}Nb_{18}W_{16}O_{93}$	1.1×10^{-13}	2.0×10^{-12}	0.27 ± 0.03
$Li_{6.8}Nb_{18}W_{16}O_{93}$	1.1×10^{-13}	1.8×10^{-12}	0.30 ± 0.01
$Li_{10.2}Nb_{18}W_{16}O_{93}$	1.1×10^{-13}	2.7×10^{-12}	0.29 ± 0.01

The lithium diffusion coefficients, D_{Li}, were measured directly at 333–453 K and extrapolated from the activation energy, E_a, to room temperature, where the relaxation time T_2 was too short to allow direct measurement. Error estimates for activation energies are derived from the standard error of the linear fit. Error bars for the diffusion coefficients are given in Extended Data Fig. 2. The two diffusion components observed in $Li_{6.3}Nb_{16}W_5O_{55}$ are denoted as '-a' and '-b' with 15% and 85% signal contribution, respectively.

Redox activity and local atomic structure

To understand (i) the nature of the charge-transfer sequence as a function of lithiation and (ii) the origin of multielectron redox behaviour in $Nb_{16}W_5O_{55}$ and $Nb_{18}W_{16}O_{93}$, the X-ray absorption near-edge structure (XANES) of the Nb K edge and W L edges was analysed (Supplementary Fig. 11, Methods). For $Nb_{16}W_5O_{55}$, operando and ex situ Nb K edge XANES spectra show a nearly linear trend between the number of electrons (and Li$^+$) transferred and the oxidation state of niobium, extracted from the shift of the absorption edge (Fig. 3a, Extended Data Fig. 3, Supplementary Fig. 12). Similarly, ex situ samples measured at the W $L_{II,I}$ edges show a steadily negative correlation between capacity and edge position; however, there appears to be a larger shift in the tungsten absorption edge for the first 0.5 Li$^+$/TM inserted (Fig. 3b, Extended Data Fig. 3, Supplementary Fig. 12), indicating a slight preference for tungsten reduction initially. The situation for $Nb_{18}W_{16}O_{93}$ (Fig. 3, Extended Data Fig. 4, Supplementary Fig. 12) is analogous in that the niobium oxidation state decreases nearly linearly and tungsten reduction is slightly favoured initially. However, there is a

© 2018 Springer Nature Limited. All rights reserved.

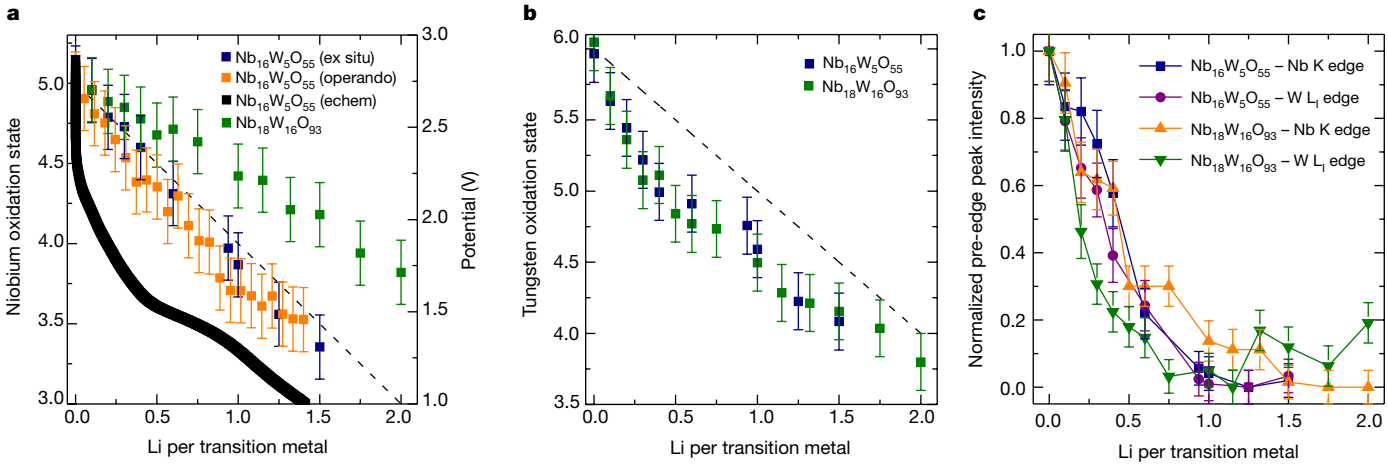


Fig. 3 | X-ray absorption spectroscopy of $Nb_{16}W_5O_{55}$ and $Nb_{18}W_{16}O_{93}$. **a**, **b**, Oxidation states of niobium (**a**) and tungsten (**b**) as a function of lithiation, extracted from shifts in the absorption energy of the Nb K edge and the W L_{II} edge. The black dashed line demarcates a slope of −1, for reference. Oxidation-state error bars are estimated to be ±0.2 on the basis of energy data resolution. The operando electrochemical discharge profile ('echem') is included in **a**. **c**, Pre-edge integrated peak intensity from the Nb K edge and W L_I edge serves as a measure of local distortion from SOJT effects on d^0 octahedral sites and, as the SOJT effect is relaxed upon reduction to d^1 (see Methods for further discussion), a further indication of the oxidation state. Pre-edge integrated peak intensity error bars represent 2σ from Levenberg–Marquardt nonlinear least-squares minimization fits (see Methods and Supplementary Fig. 13).

clear preference for multielectron reduction on the tungsten throughout the lithiation, with Nb^{3+} appearing only beyond 1.5 Li^+/TM, which may be related to cation occupancy variation on octahedral versus pentagonal bipyramidal sites.

The redox centres Nb^{5+} and W^{6+} in $Nb_{16}W_5O_{55}$ are both d^0 and both of these cations located in six-coordinate environments experience second-order Jahn–Teller (SOJT) distortions, which give rise to a pre-edge feature before the main absorption edge in the Nb K and W L_I edge X-ray absorption spectra (Extended Data Figs. 3, 4, Supplementary Figs. 13, 14, Methods); this serves as a direct probe of local symmetry and an additional measure of the oxidation state. As the d^0 cations are reduced, the energy of the d states moves up and the SOJT distortion is reduced (Methods), which increases the local octahedral symmetry and decreases the pre-edge states and intensity monotonically (Fig. 3c, Extended Data Figs. 3, 4, Supplementary Fig. 13)[23], again with a slightly larger effect on tungsten at low lithium concentrations and a bigger difference between tungsten and niobium in $Nb_{18}W_{16}O_{93}$. By about 0.8–1.0 Li^+/TM, both the Nb K and W L_I distinct pre-edges of both the block and bronze phase have decreased and no substantial further changes are observed in compositions corresponding to multi-redox; thus, lithiation is associated with an increase in local symmetry for the d^0 oxide intercalation hosts. The XANES edge shifts and pre-edge evolution for $Nb_{16}W_5O_{55}$ and $Nb_{18}W_{16}O_{93}$ observed in this work indicate a parallel reduction pathway and multielectron charge storage of both transition metals.

Anisotropic host–lattice response to Li intercalation

To probe the lattice evolution of these complex oxides upon lithiation, operando synchrotron X-ray diffraction was performed at different cycling rates. At C/2, $Nb_{16}W_5O_{55}$ evolves through a complex, three-stage solid-solution mechanism (Fig. 4a, c, Extended Data Figs. 5, 6) that correlates with the following observed electrochemical regions: (i) high voltage (until about 65 mA h g^{-1} or 0.4 Li^+/TM): a–c-plane expansion of the blocks, along with a slight expansion perpendicular to the block plane; (ii) about 65–170 mA h g^{-1} (0.4–1.0 Li^+/TM): anisotropic behaviour involving a contraction of the blocks and a substantial expansion in the (perpendicular) b direction; (iii) multi-redox (beyond 1.0 Li^+/TM): linear expansion in all dimensions. The volume expansion is considerably buffered by the block contraction in the second stage, and the lattice undergoes only 5.5% expansion at lithiation to 1.0 Li^+/TM. The diffraction results show that although the second electrochemical process has a relatively small gradient it is not two-phase, in agreement with the GITT measurements (Extended Data Fig. 1b, c). Upon charging, the stages are reversed (Extended Data Fig. 6k, l), although there is some first-cycle capacity loss. This capacity loss is ascribed, at least in part, to residual Li in the structure (Extended Data Fig. 7), rather than the usual solid–electrolyte interphase formation, because the final lithium ions are considerably harder to extract as their removal would lead to the formation of insulating domains. At a ten-times-higher rate (5C), only the first two stages of lattice evolution are observed (Extended Data Figs. 5, 6), which is consistent with the degree of loss of capacity at higher rates. In addition, there is more strain and reaction inhomogeneity at 5C, consistent with previous in situ studies that clearly showed that at these rates, lithium transport within the electrolyte results in internal (electrolyte) inhomogeneities and concentration gradients[33,34]. Nevertheless, the mechanism remains solid-solution in nature, which is evident from the high inter-peak intensity (Extended Data Fig. 6j).

Operando diffraction measurements of $Nb_{18}W_{16}O_{93}$ were performed at rates of up to 10C (Fig. 4b, d; Extended Data Figs. 5, 8, Supplementary Fig. 15). Like the block phase, the bronze phase shows a complicated but reversible nonlinear and strongly anisotropic structural evolution upon lithiation (Fig. 4b, d; see Methods for further discussion). Interestingly, the $b \approx 3a$ pseudo-superstructure relationship persists until the fourth intercalation region, at >0.5 Li^+/TM. From this point, the unit-cell volume actually decreases with increasing lithium intercalation and is the same at 1.0 Li^+/TM as at 0.5 Li^+/TM—a total increase of only 2.8% from the unlithiated host. This volume contraction upon lithium insertion is phenomenologically related to negative thermal expansion[35,36] or negative linear compressibility[37] and may have shared origins related to tilting of the polyhedral units (see 'Host features and design considerations'). The small volume change has implications for the suppression of intergranular cracking and long-term cycle performance[38]. Unlike in the shear structure, the bronze phase exhibits a compositionally narrow two-phase reaction between approximately $Li_{6.6}Nb_{18}W_{16}O_{93}$ and $Li_{10.2}Nb_{18}W_{16}O_{93}$ (0.2–0.3 Li^+/TM); the phase transition preserves the *Pbam* space-group symmetry and is characterized by a–b-plane expansion and contraction of the bronze layers. The structural stability of both the block and bronze phases is also reflected in the stability of the energy storage capacity over 1,000 discharge–charge cycles (Fig. 2f).

Host features and design considerations

Relative to the parent ReO_3 structure, the niobium tungsten oxides accommodate anion-deficient nonstoichiometry by forming topologically distinct condensed phases. This has important consequences for Li motion because the intersecting crystallographic shear planes

© 2018 Springer Nature Limited. All rights reserved.

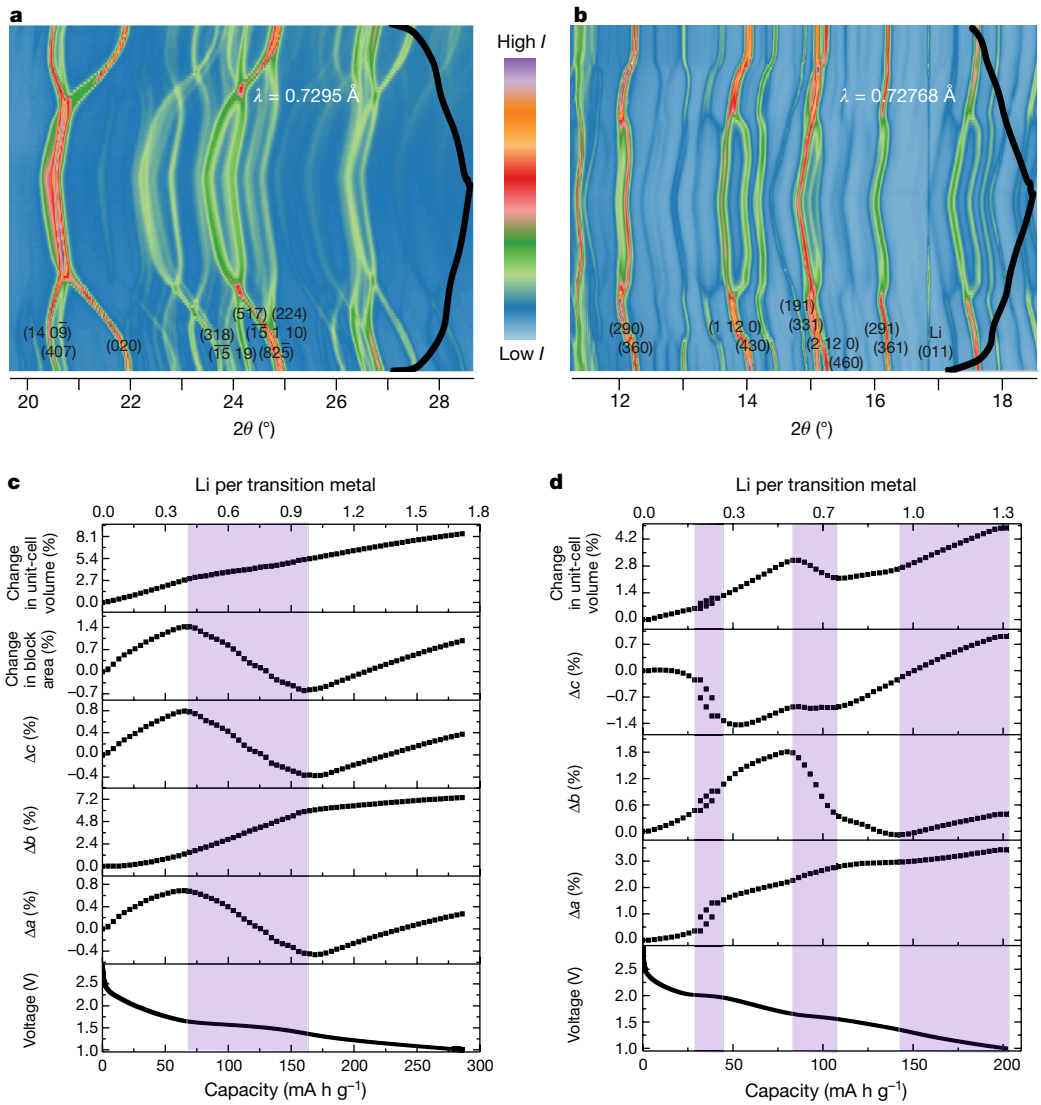


Fig. 4 | Nonlinear structural evolution of $Nb_{16}W_5O_{55}$ and $Nb_{18}W_{16}O_{93}$ from operando synchrotron diffraction. a, **b**, Operando diffraction pattern sections of $Nb_{16}W_5O_{55}$ (**a**) and $Nb_{18}W_{16}O_{93}$ (**b**), with some of the strongest reflections labelled and the electrochemistry overlaid. θ, diffraction angle; I, intensity. **c**, Upon lithiation, block-phase $Nb_{16}W_5O_{55}$ evolves in three stages in the range 3.0–1.0 V at C/2; the expansion is buffered by block contraction in the second stage. **d**, Bronze-like-phase $Nb_{18}W_{16}O_{93}$ evolves through six stages upon lithiation in the range 3.0–1.0 V at 1C. In this complex oxide, volume expansion is buffered by layer contraction in the second stage (two-phase) and b-axis contraction in the fourth and fifth stages. Alternating white and purple shading serves as a guide to the eye to demarcate reaction stages.

(block phase) or twisted octahedra locked to pentagonal columns (bronze-like) (Fig. 1a, d) decrease the structural degrees of freedom. Both the block ($Nb_{16}W_5O_{55}$) and bronze ($Nb_{18}W_{16}O_{93}$) structural motifs are derived from the parent ReO_3 structure type, of which WO_3 is a locally distorted analogue and LLTO is a version with Li^+ and La^{3+} cations on the A site (Supplementary Fig. 16). In ReO_3, WO_3 and LLTO, Li ions occupy 4- and 5-coordinate sites and diffuse through square planar window transition states. It is well established that this process is fast; an activation energy of 120–220 meV is found in LLTO, whose framework is stabilized by the large lanthanum ions[39–45]. By contrast, although ReO_3 and WO_3 should also be suitable for rapid three-dimensional lithium motion, their open and flexible framework can undergo a structural phase transition that locks the Li^+ in place, reducing Li mobility considerably[40].

The shear and bronze structure formation mechanisms result in frustrated polyhedral networks that can no longer undergo tilts upon lithiation that would clamp the lithium ions, preventing their motion. These concepts are well established in perovskites, where the so-called rigid-unit modes drive long-range cooperative distortions despite the framework structure[46,47]. By contrast, crystallographic shear and twisted, locked octahedra (Supplementary Fig. 16) lead to two fundamentally different—but effectively three-dimensional—conduction networks for lithium, while stabilizing the framework and precluding substantial Li-induced rigid-unit mode displacements that lead to local and long-range structural rearrangements.

Structural analyses and bond-valence energy landscape calculations (Extended Data Fig. 9a, b) indicate that infinite lithium diffusion in the $Nb_{16}W_5O_{55}$ block phase is one-dimensional along the b axis, but the twelve parallel tunnels act as a metaphorical multi-lane highway, enabling lithium to change 'lanes' via local hops in the a–c plane. Lithium sites and diffusion in this twelve-channel path may be analogous to those of the ReO_3 and LLTO structure types (Extended Data Fig. 9c), but without the phase transition of ReO_3 or the blocking La^{3+} ions of LLTO. Furthermore, given this mechanism, $Nb_{16}W_5O_{55}$ should not be susceptible to the tunnel-blocking defects that hinder micrometre-sized one-dimensional conductors, such as $LiFePO_4$[48–50]. The $Nb_{18}W_{16}O_{93}$ bronze-like phase has an infinite two-dimensional lithium pathway similarly to high-rate T-Nb_2O_5 in the a–b crystallographic plane, with the additional benefit of four- and five-membered

© 2018 Springer Nature Limited. All rights reserved.

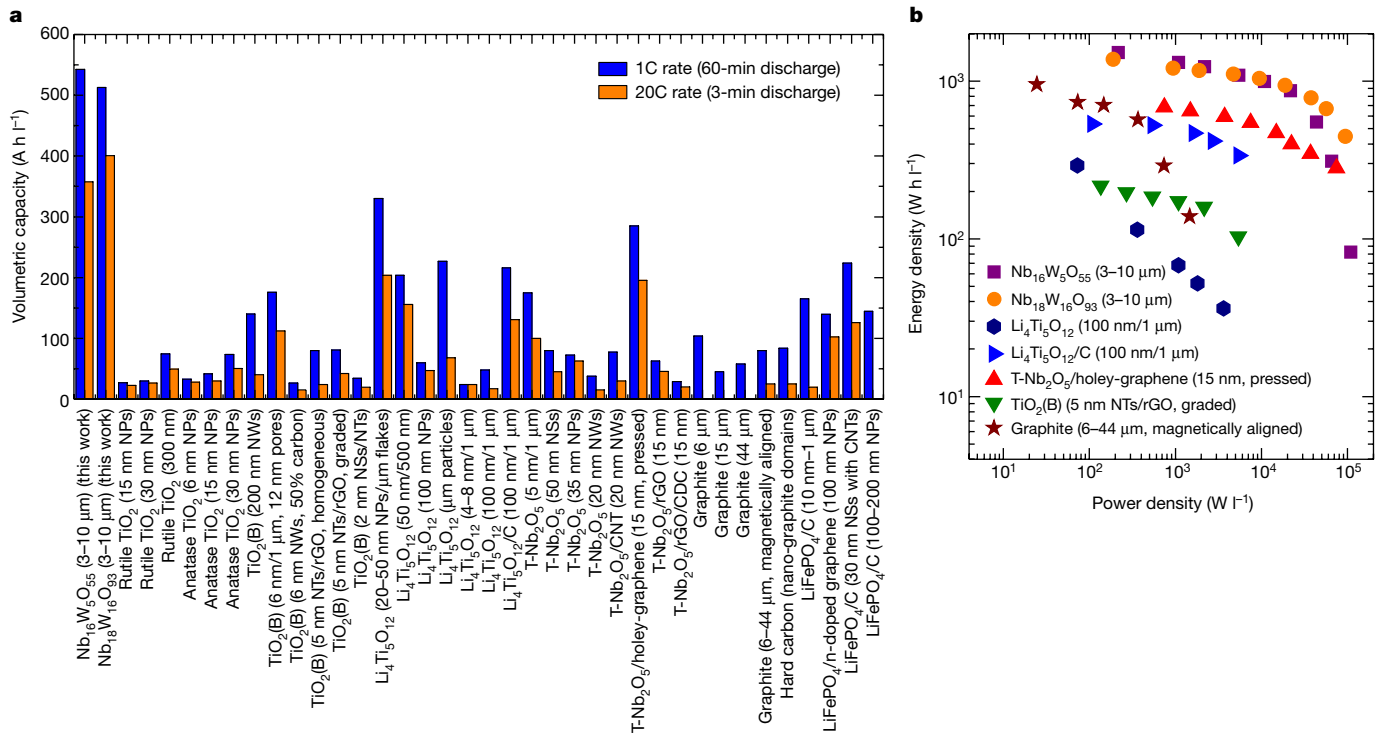


Fig. 5 | Charge, energy, and power densities of micrometre-sized $Nb_{16}W_5O_{55}$ and $Nb_{18}W_{16}O_{93}$ compared to state-of-the-art high-rate nanoscale materials and formulations. a, $Nb_{16}W_5O_{55}$ and $Nb_{18}W_{16}O_{93}$ charge-storage results from this work are compared to high-rate electrode formulations from the literature (see Supplementary Table 2). Volumetric capacities at 1C and 20C are determined from reported capacities and from reported, measured or estimated tap densities (Supplementary Table 2). The sizes of the primary and, where appropriate, secondary particles or pores are indicated. NP, nanoparticle; NW, nanowire; rGO, reduced graphene oxide; NS, nanosheet; NT, nanotube; CDC, carbide-derived carbon; CNT, carbon nanotube. **b**, Ragone (double logarithmic) plot of energy density and power density for anode active materials versus a 4.0-V cathode (Methods and Supplementary Table 2, including relevant literature references). The mass loading of the niobium tungsten oxides was 2.6 mg cm^{-2} and that of the other titanium- and niobium-based materials was about 1 mg cm^{-2}. Graphite is included as a reference, although it suffers from Li plating risks and particle fracture when used at high rates.

windows and channels along the *c* axis to increase the dimensionality of long-range diffusion to three dimensions. As a result, the structure appears to exhibit better performance, certainly under comparable morphologies and conditions, than the recently intensely studied T-Nb_2O_5[11,19,28,51,52]. The bronze comparisons stress the complementary benefits of the open room-and-pillar host structure found in T-Nb_2O_5[11] and $Nb_{18}W_{16}O_{93}$ and the additional perpendicular lithium diffusion channels in $Nb_{18}W_{16}O_{93}$. In addition, both $Nb_{16}W_5O_{55}$ and $Nb_{18}W_{16}O_{93}$ may benefit from the randomly distributed Nb and W occupancy throughout the metal sites, which prevents rate-inhibiting long-range lithium ordering(s). $Nb_{16}W_5O_{55}$ and $Nb_{18}W_{16}O_{93}$ are compared to their binary counterparts H- and T-Nb_2O_5, respectively, as a function of capacity and Li^+/TM (Extended Data Fig. 9d–g). Although the niobium tungsten oxide hosts are d^0 insulators, their three-dimensional –M–O– connectivity enables effective electronic pathways because the structure is n-doped upon lithiation. In addition, variations of the bronze and crystallographic shear structure types are abundant and tunable for specific applications and beyond Li electrochemical energy storage (see Methods for further discussion regarding structure family extensions).

Volumetric energy density and future applications

When compared strictly on the basis of the theoretical 1.0 Li^+/TM reaction and crystallographic density of the active material, titania, niobia and graphite could all display charge densities of greater than 800 A h l^{-1} (Supplementary Fig. 17). Once experimental capacities, multielectron redox and tap density are considered (Fig. 5a, Supplementary Table 2), the bulk, unoptimized niobium tungsten oxides presented here maintain volumetric charge storage densities greater than 500 A h l^{-1} at 1C and up to 400 A h l^{-1} at 20C—performance that even the most optimized versions of TiO_2, Nb_2O_5 and $Li_4Ti_5O_{12}$ cannot achieve. This does not mean that the compounds presented here cannot be improved by methods such as nanostructuring, calendaring or carbon coating (see, for example, ref. [52]), but proves that large, micrometre-sized particles can be used for high-rate electrodes and illustrates that nanoscaling is not always the most appropriate strategy to improve performance. This is evident in a Ragone plot (Fig. 5b), which shows the higher energy and power densities of the new bulk niobium tungsten oxides compared with literature values for state-of-the-art high-rate anode materials and with graphite (see expanded dataset including other electrode materials, details and references in Supplementary Table 2).

The ability to intercalate lithium into microscopic particles in minutes, with transport numbers approaching the best liquid and solid electrolytes, calls for a different paradigm for electrode structuring. Rather than focusing on particle dimensions, other aspects—such as electrolyte transport or counter electrode diffusivity and overpotential—may become more critical to enhancing performance. The strict requirements for carbon coating and intricately wiring nanoparticles are relaxed and issues stemming from surface reactivity and stability are also diminished. Other applications of these materials can be envisaged. For example, in the field of all-solid-state batteries, there has been a mismatch between lithium transport in superionic solid electrolytes compared with electrode materials. If both components have similar diffusivities, it may be possible to design and implement new, simplified composite structures to speed up the realization of safe all-solid-state energy storage.

In conclusion, extremely high-rate performance has been achieved without nanoscaling by identifying appropriate three-dimensional oxide crystal structures. The two new electrode materials, $Nb_{16}W_5O_{55}$ and $Nb_{18}W_{16}O_{93}$, effectively use superstructure motifs to provide stable host structures for lithium intercalation with facile and defect-tolerant

© 2018 Springer Nature Limited. All rights reserved.

lithium diffusion and multielectron redox. Volume expansion is mitigated by structural contraction along specific crystallographic axes in response to increased lithium content, which may enable the extended cycling of these large particles[53]. The materials investigated here operate in a similar voltage region to the well-studied, and generally considered to be 'safe', anode materials $Li_4Ti_5O_{12}$ and TiO_2(B). In contrast to approaches that try to overcome physical properties (such as ionic and electronic conductivity) extrinsically (for example, by nanoscaling and carbon coating), the high-rate, high-capacity properties of the niobium tungsten oxide block and bronze phases presented here are intrinsic to the complex atomic and electronic networks. The path forward for new fast ionic conductors should consider host structures with open yet frustrated topologies (that prevent structural rearrangements that reduce Li transport) and that also contain 'disorder' in the sense of a multitude of guest sites for Li^+ and limited interaction between the host and guest (that is, no strong coupling between the diffusing Li^+ and associated electron, as found in $LiFePO_4$[54], or between the Li^+ and the host structure itself). This leads to a relatively flat potential energy surface with small kinetic diffusion barriers for Li transport. These criteria, which lead to extremely high power and energy densities (Fig. 5b), are satisfied in $Nb_{16}W_5O_{55}$ and $Nb_{18}W_{16}O_{93}$ via crystallographic shear and pentagonal columns that prop open the framework structures and through transition-metal disorder that prevents Li ordering.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0347-0.

Received: 26 October 2017; Accepted: 11 June 2018;
Published online 25 July 2018.

1. Zhao, K., Pharr, M., Vlassak, J. J. & Suo, Z. Fracture of electrodes in lithium-ion batteries caused by fast charging. *J. Appl. Phys.* **108**, 073517 (2010).
2. Vetter, J. et al. Ageing mechanisms in lithium-ion batteries. *J. Power Sources* **147**, 269–281 (2005).
3. Downie, L. E. et al. In situ detection of lithium plating on graphite electrodes by electrochemical calorimetry. *J. Electrochem. Soc.* **160**, A588–A594 (2013).
4. Kim, C., Norberg, N. S., Alexander, C. T., Kostecki, R. & Cabana, J. Mechanism of phase propagation during lithiation in carbon-free $Li_4Ti_5O_{12}$ battery electrodes. *Adv. Funct. Mater.* **23**, 1214–1222 (2013).
5. Wang, C. et al. A robust strategy for crafting monodisperse $Li_4Ti_5O_{12}$ nanospheres as superior rate anode for lithium ion batteries. *Nano Energy* **21**, 133–144 (2016).
6. Odziomek, M. et al. Hierarchically structured lithium titanate for ultrafast charging in long-life high capacity batteries. *Nat. Commun.* **8**, 15636 (2017).
7. Oszajca, M. F., Bodnarchuk, M. I. & Kovalenko, M. V. Precisely engineered colloidal nanoparticles and nanocrystals for Li-ion and Na-ion batteries: model systems or practical solutions? *Chem. Mater.* **26**, 5422–5432 (2014).
8. Palacín, M. R., Simon, P. & Tarascon, J. M. Nanomaterials for electrochemical energy storage: the good and the bad. *Acta Chim. Slov.* **63**, 417–423 (2016).
9. Wu, H. et al. Stable cycling of double-walled silicon nanotube battery anodes through solid-electrolyte interphase control. *Nat. Nanotechnol.* **7**, 310–315 (2012).
10. Kasnatscheew, J. et al. A tutorial into practical capacity and mass balancing of lithium ion batteries. *J. Electrochem. Soc.* **164**, A2479–A2486 (2017).
11. Griffith, K. J., Forse, A. C., Griffin, J. M. & Grey, C. P. High-rate intercalation without nanostructuring in metastable Nb_2O_5 bronze phases. *J. Am. Chem. Soc.* **138**, 8888–8899 (2016).
12. Stramare, S., Thangadurai, V. & Weppner, W. Lithium lanthanum titanates: a review. *Chem. Mater.* **15**, 3974–3990 (2003).
13. Shen, L., Zhang, X., Uchaker, E., Yuan, C. & Cao, G. $Li_4Ti_5O_{12}$ nanoparticles embedded in a mesoporous carbon matrix as a superior anode material for high rate lithium ion batteries. *Adv. Energy Mater.* **2**, 691–698 (2012).
14. Prakash, A. S. et al. Solution-combustion synthesized nanocrystalline $Li_4Ti_5O_{12}$ as high-rate performance Li-ion battery anode. *Chem. Mater.* **22**, 2857–2863 (2010).
15. Xu, G. B. et al. Highly-crystalline ultrathin $Li_4Ti_5O_{12}$ nanosheets decorated with silver nanocrystals as a high-performance anode material for lithium ion batteries. *J. Power Sources* **276**, 247–254 (2015).
16. Ren, Y. et al. Nanoparticulate TiO_2(B): an anode for lithium-ion batteries. *Angew. Chem. Int. Ed.* **51**, 2164–2167 (2012).
17. Liu, H. et al. Mesoporous TiO_2–B microspheres with superior rate performance for lithium ion batteries. *Adv. Mater.* **23**, 3450–3454 (2011).
18. Beuvier, T. et al. TiO_2(B) nanoribbons as negative electrode material for lithium ion batteries with high rate performance. *Inorg. Chem.* **49**, 8457–8464 (2010).
19. Augustyn, V. et al. High-rate electrochemical energy storage through Li^+ intercalation pseudocapacitance. *Nat. Mater.* **12**, 518–522 (2013).
20. Liu, X. et al. Urchin-like hierarchical H-Nb_2O_5 microspheres: synthesis, formation mechanism and their applications in lithium ion batteries. *Dalton Trans.* **46**, 10935–10940 (2017).
21. Lai, C.-H. et al. Designing pseudocapacitance for Nb_2O_5/carbide-derived carbon electrodes and hybrid devices. *Langmuir* **33**, 9407–9415 (2017).
22. Roth, R. S. & Waring, J. L. Phase equilibria as related to crystal structure in the system niobium pentoxide–tungsten trioxide. *J. Res. Natl Bur. Stand.* **70A**, 281–303 (1966).
23. Roth, R. S. & Wadsley, A. D. Multiple phase formation in the binary system Nb_2O_5–WO_3. II. The structure of the monoclinic phases $WNb_{12}O_{33}$ and $W_5Nb_{16}O_{55}$. *Acta Crystallogr.* **19**, 32–38 (1965).
24. Roth, R. S. & Wadsley, A. D. Multiple phase formation in the binary system Nb_2O_5–WO_3. I. Preparation and identification of phases. *Acta Crystallogr.* **19**, 26–32 (1965).
25. Stephenson, N. C. A structural investigation of some stable phases in the region Nb_2O_5·WO_3–WO_3. *Acta Crystallogr. B* **24**, 637–653 (1968).
26. Naoi, K., Ishimoto, S., Isobe, Y. & Aoyagi, S. High-rate nano-crystalline $Li_4Ti_5O_{12}$ attached on carbon nano-fibers for hybrid supercapacitors. *J. Power Sources* **195**, 6250–6254 (2010).
27. Cava, R. J., Murphy, D. W. & Zahurak, S. M. Lithium insertion in Wadsley–Roth phases based on niobium oxide. *J. Electrochem. Soc.* **130**, 2345–2351 (1983).
28. Kumagai, N., Koishikawa, Y., Komaba, S. & Koshiba, N. Thermodynamics and kinetics of lithium intercalation into Nb_2O_5 electrodes for a 2 V rechargeable lithium battery. *J. Electrochem. Soc.* **146**, 3203–3210 (1999).
29. Patoux, S., Dolle, M., Rousse, G. & Masquelier, C. A Reversible lithium intercalation process in an ReO_3 type structure PNb_9O_{25}. *J. Electrochem. Soc.* **149**, A391–A400 (2002).
30. Han, J.-T., Huang, Y.-H. & Goodenough, J. B. New anode framework for rechargeable lithium batteries. *Chem. Mater.* **23**, 2027–2029 (2011).
31. Saritha, D., Pralong, V., Varadaraju, U. V. & Raveau, B. Electrochemical Li insertion studies on $WNb_{12}O_{33}$—a shear ReO_3 type structure. *J. Solid State Chem.* **183**, 988–993 (2010).
32. Griffith, K. J., Senyshyn, A. & Grey, C. P. Structural stability from crystallographic shear in TiO_2–Nb_2O_5 phases: cation ordering and lithiation behavior of $TiNb_{24}O_{62}$. *Inorg. Chem.* **56**, 4002–4010 (2017).
33. Roberts, M. R. et al. Direct observation of active material concentration gradients and crystallinity breakdown in $LiFePO_4$ electrodes during charge/discharge cycling of lithium batteries. *J. Phys. Chem. C* **118**, 6548–6557 (2014).
34. Strobridge, F. C. et al. Unraveling the complex delithiation mechanisms of olivine-type cathode materials, $LiFe_xCo_{1-x}PO_4$. *Chem. Mater.* **28**, 3676–3690 (2016).
35. Mary, T. A., Evans, J. S. O., Vogt, T. & Sleight, A. W. Negative thermal expansion from 0.3 to 1050 Kelvin in ZrW_2O_8. *Science* **272**, 90–92 (1996).
36. Lin, K. et al. Ordered structure and thermal expansion in tungsten bronze $Pb_2K_{0.5}Li_{0.5}Nb_5O_{15}$. *Inorg. Chem.* **53**, 9174–9180 (2014).
37. Cairns, A. B. & Goodwin, A. L. Negative linear compressibility. *Phys. Chem. Chem. Phys.* **17**, 20449–20465 (2015).
38. Liu, H. et al. Intergranular cracking as a major cause of long-term capacity fading of layered cathodes. *Nano Lett.* **17**, 3452–3457 (2017).
39. Murphy, D. W., Greenblatt, M., Cava, R. J. & Zahurak, S. M. Topotactic lithium reactions with ReO_3 related shear structures. *Solid State Ion.* **5**, 327–329 (1981).
40. Cava, R. J., Santoro, A., Murphy, D. W., Zahurak, S. M. & Roth, R. S. The structures of the lithium inserted metal oxides $Li_{0.2}ReO_3$ and $Li_{0.36}WO_3$. *J. Solid State Chem.* **50**, 121–128 (1983).
41. Gracia, L. et al. Composition dependence of the energy barrier for lithium diffusion in amorphous WO_3. *Electrochem. Solid State Lett.* **8**, J21–J23 (2005).
42. Shan, Y. J., Inaguma, Y. & Itoh, M. The effect of electrostatic potentials on lithium insertion for perovskite oxides. *Solid State Ion.* **79**, 245–251 (1995).
43. Chen, C. & Du, J. Lithium ion diffusion mechanism in lithium lanthanum titanate solid-state electrolytes from atomistic simulations. *J. Am. Ceram. Soc.* **98**, 534–542 (2015).
44. Jay, E. E., Rushton, M. J. D., Chroneos, A., Grimes, R. W. & Kilner, J. A. Genetics of superionic conductivity in lithium lanthanum titanates. *Phys. Chem. Chem. Phys.* **17**, 178–183 (2015).
45. Emery, J., Buzare, J. Y., Bohnke, O. & Fourquet, J. L. Lithium-7 NMR and ionic conductivity studies of lanthanum lithium titanate electrolytes. *Solid State Ion.* **99**, 41–51 (1997).
46. Giddy, A. P., Dove, M. T., Pawley, G. S. & Heine, V. The determination of rigid-unit modes as potential soft modes for displacive phase transitions in framework crystal structures. *Acta Crystallogr. A* **49**, 697–703 (1993).
47. Dove, M. T., Trachenko, K. O., Tucker, M. G. & Keen, D. A. Rigid unit modes in framework structures: theory, experiment and applications. *Rev. Mineral. Geochem.* **39**, 1–33 (2000).
48. Islam, M. S., Driscoll, D. J., Fisher, C. A. J. & Slater, P. R. Atomic-scale investigation of defects, dopants, and lithium transport in the $LiFePO_4$ olivine-type battery material. *Chem. Mater.* **17**, 5085–5092 (2005).
49. Dathar, G. K. P., Sheppard, D., Stevenson, K. J. & Henkelman, G. Calculations of Li-ion diffusion in olivine phosphates. *Chem. Mater.* **23**, 4032–4037 (2011).
50. Liu, H. et al. Effects of antisite defects on Li diffusion in $LiFePO_4$ revealed by Li isotope exchange. *J. Phys. Chem. C* **121**, 12025–12036 (2017).
51. Zhang, C. et al. Synthesis and charge storage properties of hierarchical niobium pentoxide/carbon/niobium carbide (MXene) hybrid materials. *Chem. Mater.* **28**, 3937–3943 (2016).

© 2018 Springer Nature Limited. All rights reserved.

52. Sun, H. et al. Three-dimensional holey-graphene/niobia composite architectures for ultrahigh-rate energy storage. *Science* **356**, 599–604 (2017).
53. Zhang, S. Chemomechanical modeling of lithiation-induced failure in high-volume-change electrode materials for lithium ion batteries. *Npj Comput. Mater.* **3**, 7 (2017).
54. Maxisch, T., Zhou, F. & Ceder, G. Ab initio study of the migration of small polarons in olivine Li_xFePO_4 and their association with lithium ions and vacancies. *Phys. Rev. B* **73**, 104301 (2006).

Acknowledgements K.J.G. acknowledges support from The Winston Churchill Foundation of the United States, a Herchel Smith Scholarship and a Science and Technology Facilities Council Futures Early Career Award. K.J.G. and C.P.G. thank the EPSRC for a LIBATT grant (EP/M009521/1). L.E.M. acknowledges support from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska–Curie grant agreement number 750294 and a Charles and Katharine Darwin Research Fellowship. We thank I. Seymour from the University of Cambridge and B. Dunn from the University of California, Los Angeles for fruitful discussions. We thank J. Skepper and H. Greer from the University of Cambridge for assistance with electron microscopy and M. Avdeev from the Bragg Institute for the bond valence sum mapping program. We thank O. Borkiewicz from the Advanced Photon Source at Argonne National Laboratory and A. Kasam from the University of Cambridge for diffraction data reduction scripts. We thank Diamond Light Source for access to beamline B18 (SP11433, SP14956, SP16387, SP17913), where we obtained results presented here. This research used resources of the Advanced Photon Source (GUP40466, GUP41657, GUP47967), a US Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under contract number DE-AC02-06CH11357.

Reviewer information *Nature* thanks S. Greenbaum, P. Woodward and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions K.J.G. and C.P.G. conceived the idea. K.J.G. synthesized and characterized the materials and performed the electrochemistry experiments. K.J.G. performed the synchrotron diffraction and absorption experiments and analysed the data with support from K.M.W. and G.C. L.E.M. and K.J.G. performed the PFG NMR measurements. K.J.G. and C.P.G. wrote the manuscript with input from all co-authors.

Competing interests K.J.G. and C.P.G., via Cambridge Enterprise, have filed a UK patent application (GB1809467.2) covering the materials and high-rate energy storage application described in this manuscript.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0347-0.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0347-0.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to C.P.G.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Synthesis and structure characterization. $Nb_{16}W_5O_{55}$ and $Nb_{18}W_{16}O_{93}$ were synthesized by co-thermal oxidation of dark-blue NbO_2 (Alfa Aesar, 99+%) and brown WO_2 (Alfa Aesar, 99.9%) in batches of approximately 1–5 g. The partially reduced oxides were massed to within 0.001 g of the 16:5 or 18:16 stoichiometric ratios, ground together by hand with an agate mortar and pestle, pressed into a pellet at 10 MPa, heated in a platinum crucible at a rate of 10 K min^{-1} to 1,473 K, and naturally cooled in the furnace over approximately 2 h. Both phases formed subhedral particles with approximately 3–10-μm primary particles that appeared in the electron microscope to be intergrown/'cemented' into larger secondary particles of around 10–30 μm (Fig. 1, Supplementary Fig. 3a, b). As synthesized, $Nb_{16}W_5O_{55}$ is pale yellow–green and $Nb_{18}W_{16}O_{93}$ is off-white (Supplementary Fig. 3c) (we note that the colours of both mixed metal oxides are between those of light-green WO_3 and white Nb_2O_5). Energy-dispersive X-ray spectroscopy (EDX) revealed a uniform distribution of niobium, tungsten and oxygen, with no apparent phase-segregated particles and no presence of any other elements (besides those on the background adhesive carbon) (Supplementary Figs. 4, 5). We note that although phases such as Nb_2O_5 and WO_3 would be distinct via EDX, particles of niobium tungsten oxide phases with similar composition to that of $Nb_{16}W_5O_{55}$ or $Nb_{18}W_{16}O_{93}$ would not be distinguishable. Although the phase diagram[55] suggests that $Nb_{16}W_5O_{55}$ is only formed at about 1,363–1,658 K and is a metastable phase that should disproportionate to $Nb_{14}W_3O_{44}$ and Nb_2WO_8 on cooling, this was not observed. $Nb_{16}W_5O_{55}$ retains its high-temperature structure on quenching[56] and even with more modest cooling. In the original synthetic report[24], Roth and Wadsley also noted that phase-pure $Nb_{16}W_5O_{55}$ was obtained even after annealing at 1,273 K for one to three days after an initial heat treatment at 1,623 K for 24 h. To investigate other synthetic methods, samples were also prepared by quenching the samples from 1,473 K onto a steel plate, as well as from analogous pellets or loose mixed powders of Nb_2O_5 (Alfa Aesar, 99.9985%, H-phase) and WO_3 (Sigma-Aldrich, ≥99%, γ-phase) precursors. Large (20 g) single batches of $Nb_{16}W_5O_{55}$ and $Nb_{18}W_{16}O_{93}$ were produced by hand grinding (agate, 5 min) or ball milling (SPEX 8000M high-energy mixer/mill; zirconia jar with two 5-g zirconia balls, 90 min) Nb_2O_5 and WO_3 powder, followed by heating in a platinum crucible covered with a platinum lid in air at 973 K for 12 h and then at 1,473 K for 12 h. The large batches were heated at 10 K min^{-1} and the samples were allowed to cool naturally in the furnace over about 2 h. Thermogravimetric analysis (TGA, Supplementary Fig. 6) revealed that WO_2 oxidizes to WO_3 starting at around 700 K and NbO_2 oxidizes to Nb_2O_5 starting at around 550 K. At 1,273 K, the WO_2 sample had gained 7.8% mass versus the expected 7.4%, and NbO_2 had gained 5.8% compared to the expected 6.4%. These small differences may arise from slight off-stoichiometry in the starting material. WO_3 undergoes sublimation at temperatures near or above 1,273 K; however, the reaction with niobium oxide apparently stabilizes the volatile tungsten to considerably higher temperatures. After 24 h at 1,473 K, the mass change was –0.4% for $Nb_{16}W_5O_{55}$ and –0.3% for $Nb_{18}W_{16}O_{93}$ compared with the expected values, which is within the error of the expected mass changes for NbO_2 and WO_2 oxidation.

Laboratory powder X-ray diffraction patterns were measured at ambient temperature with a Cu Kα source in Bragg–Brentano geometry at a scan rate of 0.83° min^{-1}. The sample stage rotated continuously to improve powder averaging. Sample phase and purity were determined by Rietveld refinement in GSAS-II (Supplementary Fig. 7)[57]. Given the synthesis time and temperature (24 h and 1,473 K), the block-phase $Nb_{16}W_5O_{55}$ samples are expected to contain several per cent of Wadsley defect fringes, as suggested by the laborious study of Allpress and Roth[58].

TGA measurements were performed on a Mettler Toledo TGA/SDTA 851 thermobalance. Samples were placed in a tared 100-μl alumina crucible and the mass was recorded from 323 K to 1,273 K in steps of 1 K min^{-1} under a constant air flow (50 ml min^{-1}). A blank with an empty crucible was recorded under the same conditions and subtracted from the NbO_x/WO_x data. The raw data were numerically differentiated to obtain differential thermogravimetry curves.

Scanning electron microscopy images were taken with a Sigma VP instrument (Zeiss) at 3.0 kV and a MIRA3 instrument (TESCAN) at 5.0 kV with secondary-electron detection.

Tap density was recorded on an AutoTap (Quantachrome Instruments) instrument operating at 257 taps per minute. Tap densities were measured according to the ASTM international standard B527-15, modified to accommodate a 5–10 cm^3 graduated cylinder.

Electrochemical characterization. All electrochemical measurements were conducted in stainless-steel 2032 coin cells (Cambridge Energy Solutions) with a stainless-steel conical spring, two 0.5-mm-thick stainless-steel spacer disks, a plastic gasket and a glass microfibre separator (Whatman). The electrode had an active material:carbon:binder mass ratio of 8:1:1, active material loading of 2–3 mg cm^{-2} and an area of 1.27 cm^2. The metal oxide and conductive carbon (Super P, TIMCAL) were ground by hand in an agate mortar and pestle in an 8:1 mass ratio. This powder was ground in a 9:1 mass ratio with poly(vinylidene difluoride) (PVDF, Kynar) dispersed in *N*-methyl pyrrolidone (NMP; Sigma-Aldrich, anhydrous, 99.5%). Although standard, SuperP carbon is a nanoparticulate powder and NMP is a hazardous organic solvent, so appropriate nanoparticle cabinets and fume hoods should be used. This metal oxide/carbon/polymer electrode served as the cathode against a Li metal-disk (LTS Research, 99.95%) anode in half-cell geometry. The electrolyte for all experiments was 1.0 M $LiPF_6$ dissolved in 1:1 v/v ethylene carbonate/dimethyl carbonate (EC/DMC; Sigma-Aldrich, battery grade). No additives were used. Electrochemistry measurements were performed in a temperature-controlled room at 293 ± 2 K. A galvanostat/potentiostat (BioLogic) with EC-Laboratory software was used to perform the electrochemical measurements. All potentials are reported with respect to Li^+/Li^0.

In this work, the C-rate is always defined relative to one-electron transfer per transition metal; for example, for $Nb_{16}W_5O_{55}$, 1C = 171.3 mA g^{-1} and 20C = 3,426 mA g^{-1}. Theoretical capacity is calculated by:

$$Q_{\text{theoretical}} = \frac{nF}{3.6m} = \frac{21 \times 96{,}485.3 \text{ C mol}^{-1}}{3.6 \text{ C mA}^{-1}\text{ h}^{-1} \times 3{,}285.65 \text{ g mol}^{-1}} = 171.3 \text{ mA h g}^{-1}$$

where n is the number of electrons transferred per formula unit, F is Faraday's constant, 3.6 is a conversion factor between coulombs and the conventional milliampere-hour and m is the mass per formula unit.

In a Ragone plot[59], the energy density of a battery cathode material or full cell is the product of capacity (Q) and voltage (V); however, this quantity is not appropriate when comparing anode materials, where energy and voltage have an inverse relationship. In the calculation of the anode-material Ragone plot in Fig. 5b, the energy (E) is computed on the basis of the voltage difference versus a 4.0-V cathode. Thus, normalized to the anode, $E_{\text{anode}} = Q_{\text{anode}}(V_{\text{cathode}} - V_{\text{anode}})$. Literature values for the Ragone plot were extracted with WebPlotDigitizer[60].

Operando synchrotron X-ray diffraction. Operando powder diffraction measurements were carried out at beamline 17BM (bending magnet) at the Advanced Photon Source at Argonne National Laboratory. The samples were measured at ambient temperature in transmission geometry at 17.0 keV (0.7295 Å or 0.72768 Å) with an area detector. All operando measurements were performed in the AMPIX cell, which has been described elsewhere[61]. In brief, it contains a hard, conductive glassy carbon window to prevent inhomogeneous electrochemical reactions, which are a concern with flexible or non-conductive X-ray windows[61]. Owing to the absorbing nature of niobium and tungsten and the absence of a current collector, self-standing electrodes were made with a 3:6:1 ($Nb_{16}W_5O_{55}$ at C/2) or 5:4:1 ($Nb_{16}W_5O_{55}$ at 5C and $N_{18}W_{16}O_{93}$ at 1C, 5C and 10C) mass ratio of metal oxide:SuperP carbon:poly(tetrafluoroethylene). The resulting electrodes were approximately 1.0 cm in diameter and contained 5.3 mg, 8.7 mg and 8.0 mg of active material for $Nb_{16}W_5O_{55}$ (C/2), $Nb_{16}W_5O_{55}$ (5C) and $N_{18}W_{16}O_{93}$ (1C, 5C and 10C), respectively. These electrodes correspond to areal metal-oxide mass loadings of 6.8–11.1 mg cm^{-2}. AMPIX cells were constructed in an argon glovebox with lithium-metal counter-electrodes, glass-fibre separators (Whatman) and 1.0 M $LiPF_6$ dissolved in 1:1 v/v ethylene carbonate/diethyl carbonate (EC/DEC). Two-dimensional image data were converted to conventional one-dimensional diffraction patterns through integration in GSAS-II after calibration with LaB_6[57]. Background subtraction, primarily from the glassy carbon window, and normalization to incident X-ray intensity were performed on the one-dimensional integrated diffraction data.

Sequential Rietveld refinement of the unit-cell parameters of $Nb_{16}W_5O_{55}$ and $Nb_{18}W_{16}O_{93}$ during operando galvanostatic electrochemical lithiation was performed in GSAS-II. The initial structure model was based on Rietveld refinement of the atomic coordinates, unit-cell parameters, background, zero offset and peak-shape parameters (U, V and W; peak width squared $(H^2) = U\tan^2\theta + V\tan\theta + W$)[62] of the first in situ diffraction pattern before the application of a current. For the sequential refinement, only the unit-cell parameters and Chebyshev background coefficients were allowed to vary. Refinement of the atomic coordinates was not attempted owing to the large number of variables this would introduce, the weak relative scattering of lithium and oxygen versus niobium and tungsten, and the optimization of experiments towards rapid data collection. $Nb_{16}W_5O_{55}$ contains 11 metal sites and 28 oxygen sites and $Nb_{18}W_{16}O_{93}$ contains 10 metal sites and 25 oxygen sites in the asymmetric unit. Two-dimensional diffraction images (integrated to one-dimensional patterns) were collected in 7–15 s (Supplementary Table 3).

$Nb_{18}W_{16}O_{93}$, which exhibits a region of two-phase behaviour, was fitted using a double-ended sequential Rietveld refinement, in which the process described above was also initiated from the structure at the end of the discharge and a 'backward' refinement was performed. Error bars (Extended Data Fig. 5) are derived from the refinements. The unit-cell parameters in the two-phase region of $Nb_{18}W_{16}O_{93}$ and the inhomogeneous solid-solution region of $Nb_{16}W_5O_{55}$ were estimated by fitting individual reflections to two sets of lattice parameters ($Nb_{18}W_{16}O_{93}$) and measuring the range of individual reflections ($Nb_{16}W_5O_{55}$); thus, error bars in these

© 2018 Springer Nature Limited. All rights reserved.

regions cannot be calculated using the whole pattern-refinement method. Owing to the unit-cell dimensions and pseudosymmetry, there is severe overlap of reflections in the powder diffraction patterns of both $Nb_{18}W_{16}O_{93}$ and $Nb_{16}W_5O_{55}$, even at the high resolution available at a synchrotron X-ray source. Furthermore, anisotropic lattice expansion results in reflections crossing during discharge (Fig. 4), which introduces challenges for sequential refinement. At 145–155 mA h g^{-1} in $Nb_{16}W_5O_{55}$, this crossover (Fig. 4a, Extended Data Fig. 6k) resulted in strongly correlated and inconsistent fits for four consecutive diffraction patterns. These patterns were fitted according to the individual reflection model described above assuming linear changes—an approximation over the small change in lithium content and unit-cell dimensions.

For the crystallographic shear structure with monoclinic symmetry ($Nb_{16}W_5O_{55}$; Fig. 1a, Fig. 4c), the *b* axis represents the 'layered' direction and the 'block area' is parameterized by $ac\sin\beta$, where the *a* and *c* axes define the plane of the block and β is the angle between *a* and *c*. The unit-cell volume of this phase is the product of the block area and the layer spacing, $(ac\sin\beta)b$. For the bronze-like phase with orthorhombic symmetry ($Nb_{18}W_{16}O_{93}$, Fig. 1d, Fig. 4d), the unit-cell volume is simply the product of the crystallographic axes, *abc*.

Operando and ex situ synchrotron X-ray absorption spectroscopy. Operando X-ray absorption spectroscopy (XAS) was performed at beamline 9BM (bending magnet) at the Advanced Photon Source at Argonne National Laboratory. Nb K edge spectra were recorded at ambient temperature in transmission mode above and below the absorption edge of 18,986 eV. The same AMPIX cells were used as for the diffraction measurements, but with a higher metal-oxide loading of 22.5 mg and lower carbon content (8:1:1 ratio of oxide:carbon:binder) at C/3.

Ex situ XAS was performed at beamline B18 (bending magnet) at the Diamond Light Source. Spectra of the Nb K edge, W L_I, L_{II} and L_{III} edges were measured at ambient temperature in transmission mode using energy scans above and below the absorption edges of 18,986 eV, 12,100 eV, 11,544 eV and 10,207 eV, respectively. The X-ray energy ranges were scanned over 1,000 eV (W $L_I + L_{II}$) or 1,500 eV (Nb K, W L_{III}) in steps of about 0.4 eV. To synthesize powders for ex situ characterization, pure $Nb_{16}W_5O_{55}$ powder was pressed into pellets at 1–2 MPa, discharged to the relevant lithium content in coin cells as previously described, then extracted in an argon glovebox where it was rinsed three times with 2 ml DMC (Sigma-Aldrich, anhydrous, ≥99%) and hand-ground (agate) back into powder. Upon lithiation, the colour of $Nb_{16}W_5O_{55}$ changed from a pale yellow–green to light blue, then to dark blue, and through to dark maroon above approximately 0.5 Li^+/TM. Samples for XAS were prepared in an argon glovebox by carefully grinding the active material (typically 15–18 mg) with dried cellulose (to a total mass of 60 ± 5 mg) in an agate mortar and pestle to achieve a fine homogeneous sample with a change in absorption coefficient of about 1 across the Nb K and W L_{III} absorption edges. The thoroughly ground powders were pressed into 8-mm-diameter pellets (about 1 mm thick) in the glovebox. Non-air-sensitive samples (pristine and reference) were measured in air and lithiated samples were transferred from the glovebox in a custom-built (Diamond Light Source) transfer chamber with X-ray-transparent windows. The argon in the chamber was flushed with helium to reduce background absorption of the incident X-rays and the samples were measured at a slight overpressure of helium to ensure the exclusion of air. Multiple spectra were collected from each sample to improve the signal-to-noise ratio and to ensure sample stability with respect to possible air contamination or beam damage. The oxides appeared to be stable; no changes were observed for any sample between the first and last measurement (typically 2–6 spectra per sample per absorption edge) over several hours in the sample chamber and 30–60 min in the X-ray beam.

For all XAS measurements, simultaneous measurement of a reference Nb or W foil was performed with each spectrum to ensure X-ray beam stability and for energy calibration. Data merging, calibration/alignment, analysis and peak fitting were performed within the Athena program in the Demeter package running IFEFFIT[63,64]. The Nb K edge, W L_I, L_{II} and L_{III} edges were calibrated to the reference Nb and W metal foils by setting the first maximum in the derivative plot of absorption versus energy equal to the standard electron binding energy[65]. The $Nb_{16}W_5O_{55}$ operando calibration was set to be equivalent to the ex situ calibration to simultaneously compare edge shifts of both datasets.

XAS of the Nb K edge probes excitations from 1*s* to 5*p* at energies around 19,000 eV, which is well suited for standard synchrotron analysis. In principle, the same type of analysis could be performed with excitations from 1*s* to 6*p* at the W K edge, but at 69,500 eV, it is impractically high in energy for standard beamline configurations and suffers from severe core-hole lifetime broadening[66]. The W L edges, however, at 12,100–10,200 eV are suitable for high-resolution XAS. The W L_I edge measures 2*s* to 6*p* excitations whereas the W L_{II} and L_{III} edges correspond to transitions of 2*p* to (near-)valence 5*d* states. The L_{II} and L_{III} edges are split by about 1,337 eV by spin–orbit coupling; in the octahedral sites present in $Nb_{16}W_5O_{55}$, these edges are further split by about 4 eV by ligand-field effects into transitions to t_{2g} and e_g orbitals. Recent studies have demonstrated the use of the 5*d* L_I edges for oxidation-state analysis in periodic[11] and molecular[67,68] structures. We considered this; however, the edge position of the W L_I edge in $Nb_{16}W_5O_{55}$ is severely complicated by the strong pre-edge feature. In this system, the L_{II} edge offered an excellent alternative, with a stronger signal and a single contribution from excitations from $2p_{1/2}$ to $5d_{3/2}$. We note that the contribution expected for 2*p* to 6*s* transitions can be neglected at the absorption edge for early-third-row transition metals[69]. Meanwhile, the W L_I pre-edge peaks that are problematic for energy calibration provide direct evidence for local geometry around the tungsten ions.

An SOJT distortion moves Nb^{5+} and W^{6+} off-centre in the otherwise centrosymmetric environments of the octahedral MO_6 transition-metal sites in $Nb_{16}W_5O_{55}$ and $Nb_{18}W_{16}O_{93}$. The SOJT results from the vibronic mixing of non-degenerate electronic states under nuclear displacements[70]. In a d^0 metal oxide, the energy difference between the ground and excited states is small owing to the pseudodegeneracy of the filled valence *p* orbitals and empty *d* orbitals in the first excited state. This off-centre distortion enables mixing of the dipole-transition-allowed *p* states (for example, 5*p* for the Nb K edge) with lower-energy *d* states (4*d* for the Nb K edge), which gives rise to a pre-edge feature before the main absorption edge in the XAS spectra of the Nb K and W L_I edges and serves as a direct probe of local symmetry and an additional measure of the oxidation state. As the d^0 cations are reduced, the energy of the *d* states moves up and the SOJT distortion is reduced, which increases the local octahedral symmetry and decreases the pre-edge states and intensity. X-ray absorption pre-edge features are well known in tetrahedral and other noncentrosymmetric local-coordination environments, so the SOJT distortion and d^0 electronic configuration are not prerequisites for pre-edge peak intensity from those sites. $Nb_{16}W_5O_{55}$ has 40 distorted octahedral sites and 2 tetrahedral sites per unit cell whereas $Nb_{18}W_{16}O_{93}$ has 30 octahedral sites and 4 occupied pentagonal bipyramidal sites per unit cell, so some pre-edge intensity may be expected, even beyond lithiation to the d^1 electronic configuration. Interestingly, it has been shown that the tetragonal bronze structure cannot be constructed from identical, regular octahedra[71], so some local distortion is topologically required regardless of the oxidation state or SOJT.

We note that for the related shear-phase $Nb_{12}WO_{33}$, the three observed electrochemical regions have tentatively been assigned to the sequential reduction of $W^{6+/5+}$, $Nb^{5+/4+}$ and $Nb^{4+/3+}$ on discharge, but these proposed oxidation changes have not been experimentally verified[31,72].

^{7}Li PFG NMR spectroscopy. ^{7}Li NMR diffusion spectra were recorded on a Bruker Avance III 300 MHz (7.0 T) spectrometer using a Diff50 probe head equipped with an extended variable-temperature 5-mm single-tuned ^{7}Li saddle coil insert. All of the materials examined here had spin–lattice (T_1) relaxation times greater than spin–spin (T_2) relaxation times. Therefore, spectra were recorded with the stimulated-echo PFG sequence shown in Extended Data Fig. 2b to minimize T_2 losses. After the first 90° radiofrequency pulse, the net magnetization loses coherence owing to T_2 relaxation; thus, the time period following this pulse, which includes the first PFG pulse (to encode spin position), must be shorter than T_2. In the stimulated-echo sequence used here, a second 90° radiofrequency pulse was applied to store the net magnetization along the *z* axis, allowing diffusion times, Δ, commensurate with the comparatively longer T_1 value, as no T_2 relaxation occurred. During Δ, a short spoiler gradient (SINE.100) was applied to remove residual transverse magnetization. Afterwards, a third 90° radiofrequency pulse was applied, followed by a PFG pulse to decode the spin position. Sufficiently long delays (1 ms) were used between PFG and radiofrequency pulses to minimize eddy currents in the diffusion measurements.

During this sequence, the gradient strength, *g*, was varied from 0.87 to 1,800 or 2,300 G cm^{-1}, and 16 gradient steps were acquired using 'opt'-shaped pulses with 1,024–4,096 transients. The opt shape is a composite pulse that starts with a quarter of a sine wave, followed by a constant gradient, and ends with a ramp down (Extended Data Fig. 2b). The opt gradient pulses provided the largest gradient integral for a given time period, maximizing the range of diffusion coefficients that we could assess in this experiment.

Spectra were analysed in phase-sensitive mode and the response of the integrated NMR signal intensity, *I*, to variation in *g* is described by the Stejskal–Tanner equation[73]:

$$\frac{I}{I_0} = \exp\left[-g^2\gamma^2\delta^2\left(\Delta - \frac{\delta}{3}\right)D\right]$$

where I_0 is the intensity in the absence of gradients, γ is the gyromagnetic ratio ($\gamma_{^7Li} = 103.962 \times 10^6$ s^{-1} T^{-1}), δ is the effective gradient pulse duration and *D* is the diffusion coefficient. Typical δ values ranged from 0.8 ms to 1.5 ms and Δ values were 50–150 ms for the bronze- and the block-phase samples, respectively.

Diffusion spectra were recorded at elevated temperatures owing to the increase in T_2 observed at high temperature (for example, T_2 for $Li_{3.4}Nb_{18}W_{16}O_{93}$ is approximately 700 μs at room temperature and 1.9 ms at 453 K). (No attempt was made to calibrate the temperature for this experimental setup because a single-tuned ^{7}Li coil was used and no reliable ^{7}Li reference is routinely used for

© 2018 Springer Nature Limited. All rights reserved.

temperature calibration. The Bruker manual states that for static measurements, the temperature calibration should be within ±7 K of the set value.) The increase in T_2 allowed the use of longer gradient pulses, δ, that were necessary to measure diffusion coefficients in the solid oxides.

Representative ^{7}Li diffusion decay curves for the five samples studied, $Li_xNb_{16}W_5O_{55}$ ($x = 6.3$, 8.4) and $Li_xNb_{18}W_{16}O_{93}$ ($x = 3.4$, 6.8, 10.2), and a representative one-dimensional ^{7}Li NMR spectrum are shown in Extended Data Fig. 2.

Concerning the difficulties associated with measuring self-diffusion coefficients with PFG NMR spectroscopy, it is worth noting that even if the T_2 of these materials was sufficiently long, the cumulative length of δ applied before Δ in the pulse sequence of choice (Extended Data Fig. 2b) should not exceed 5 ms to mitigate probe damage. The limitation of cumulative $\delta = 5$ ms places an inherent limitation on the diffusion coefficients that can be extracted using PFG NMR spectroscopy and effectively places the burden on the length of T_1, which determines the length of Δ.

The PFG data were initially fitted with a single exponential to extract the diffusion coefficients. In the case of the block phase, $Li_{6.3}Nb_{16}W_5O_{55}$, fitting the data with two components for biexponential decay resulted in an improvement (on average) of a factor of two in the residual sum of squares of the fit to the PFG data. Therefore, the ^{7}Li NMR signal of the $Li_{6.3}Nb_{16}W_5O_{55}$ sample presented in Table 1 can be interpreted as consisting of two Li species, one that diffuses rapidly and one that diffuses more slowly (we note that both diffusion coefficients are markedly faster than those reported for $Li_4Ti_5O_{12}$ and $LiFePO_4$). For the higher-Li-content block phase, $Li_{8.4}Nb_{16}W_5O_{55}$, the data were fitted well by a single-component exponential that yielded a similar diffusion coefficient and activation energy to the majority lithium phase in $Li_{6.3}Nb_{16}W_5O_{55}$. This suggests that within the two-component model, the fastest diffusing phase is present at lower lithium concentrations, and by the second stage of the electrochemical profile the structure is composed entirely of the phase with room-temperature lithium diffusion of about 1.6×10^{-13} m^2 s^{-1}.

For all measured stages of lithiation for the bronze phase ($Li_{3.4}Nb_{18}W_{16}O_{93}$, $Li_{6.8}Nb_{18}W_{16}O_{93}$ and $Li_{10.2}Nb_{18}W_{16}O_{93}$), the data were fitted well by a single-component exponential decay (as reflected in the error bars in Extended Data Fig. 2a, which represent 2σ values obtained from the fit), and no remarkable improvement was observed upon fitting by a biexponential.

Linear fits of the ^{7}Li diffusion coefficients of both the block and bronze phases as a function of temperature allow extraction of the activation energy, assuming Arrhenius behaviour (Extended Data Fig. 2a). Extrapolation using the activation energy allows estimation of ^{7}Li diffusion at room temperature and facilitates comparison to other materials (Table 1, Extended Data Table 1).

GITT diffusion coefficient and diffusion length. Information on electrode thermodynamics, including phase transitions, and lithium kinetics[74] can in principle be extracted from GITT measurements by tracking the voltage evolution after a brief current pulse as lithium diffuses and the chemical potential equilibrates within the electrode/particles. Reliable quantitative diffusion coefficients, D_{Li}, are however difficult to extract from GITT alone. To determine a diffusion coefficient from GITT measurements, a diffusion length (L) must be defined. However, a battery electrode is a heterogeneous system. First, it is a composite of active material (here, metal oxide), porous carbon and polymeric binder. Within this composite, there is a distribution of particle sizes (unless single crystals[75] or well defined particles are used; even then, diffusion varies with lattice direction). Furthermore, different regimes of diffusion must exist because of the presence of solid/liquid interfaces and porous electrode structure. Nevertheless, in an electrode that does not undergo severe pulverization (for example, an intercalation electrode), L is a fixed quantity throughout the experiment. Variation in L—a parameter required to relate the rate of relaxation to the diffusion—causes values of D_{Li} to vary considerably between reports, even for the same material[28,76]. Thus, although a physically meaningful diffusion coefficient may not be extracted, a relative measure of diffusion is readily obtained. For this reason, we use an extracted proxy for lithium diffusion ($D_{Li}L^{-2}$; Extended Data Fig. 1), which removes the uncertainty in L and enables self-consistent analysis of a single electrode and electrodes prepared under identical conditions. The addition of quantitative information from another method (for example, NMR spectroscopy or tracer diffusion) allows us to calibrate relative changes in Li^+ kinetics to quantitative diffusion values throughout a range of lithiation.

$Nb_{18}W_{16}O_{93}$ operando diffraction mechanism. As shown in Fig. 4d, lithiation from 0 to 0.2 Li^+/TM is characterized by a–b-plane expansion. The plateau-like region from 0.2 to 0.3 Li^+/TM is associated with further a–b-plane expansion, still maintaining the $b \approx 3a$ superstructure and contraction of the layers (c axis). From 0.3 to 0.55 Li^+/TM, the structure expands nearly isotropically, whereas at about 0.55 Li^+/TM, the pseudo-superstructure relationship collapses. From this point until 0.75 Li^+/TM, b contracts rapidly, a expands and c is constant. At 0.75 Li^+/TM, the layers begin to expand back to their initial interlayer spacing while b continues contracting, albeit at a slower rate, to its initial length, and a continues its expansion. Thus, when fully lithiated to 1.0 Li^+/TM ($Li_{34}Nb_{18}W_{16}O_{93}$), b and c are within ±0.1% of the unlithiated host, while a, and thus the total volume expansion, is +2.8%. In the multi-redox region beyond 1.0 Li^+/TM, the structure again expands somewhat isotropically.

Structure family extensions. Regarding the electronic structure changes of $Nb_{16}W_5O_{55}$ and $Nb_{18}W_{16}O_{93}$ upon lithiation, related partially reduced phases may be considered. Very high electronic conductivities are known to exist in the crystallographic shear niobium tungsten oxides[77], as well as related partially reduced versions of both families, for example, shear $Nb_{12}O_{29}$[78] and bronze $A_{0.3}WO_3$[79] (A = alkali metal cation), which exhibit resistivities lower than 10^{-2} Ω cm.

For ionic diffusion, there are also analogies between the new lithiated complex oxides and superionic conductors, such as α-AgI[80] and $Li_{10}GeP_2S_{12}$[81,82], in that there are multiple sites for the cations to occupy within a disordered three-dimensional structure[83,84].

In the composition and phase space of block and bronze structures, cation (such as early transition metals, p-block elements, aliovalent mixed metals)[27,75,77,85,86] and anion (such as oxyfluorides[27,87,88]) dopants make it possible to tune the ionic and electronic properties and thereby affect voltage, capacity, power capability and stability. Analogues are known with Na^+, K^+, Mg^{2+} and Ca^{2+}, which suggests that the variety of possible tunnel shapes and sizes, vacant sites and three-dimensional connectivities is also promising for beyond-Li electrochemical energy storage, according to the insights into high rate and capacity discovered here.

Application considerations. Fast charging or high-power delivery from a full cell requires a cathode to match the anode. Although $LiFePO_4$ has been used as a promising high-rate cathode[89], along with $Li_4Ti_5O_{12}$, both of these electrodes have exceptionally flat voltage profiles. The $LiFePO_4$–$Li_4Ti_5O_{12}$ cathode–anode combination provides a constant voltage but presents a serious challenge in terms of battery management systems (BMS). Simple and accurate BMS are crucial for battery applications in electric vehicles and mobile technology and are even more important at high rates for preventing dangerous and degradative over(dis)charging while maximizing utility. BMS rely on their ability to measure the state of charge, which cannot be done by charge counting alone as the battery degrades. The open-circuit voltage is a more reliable measure and a thermodynamic quantity. In contrast to the $Li_4Ti_5O_{12}$/$LiFePO_4$ battery, the sloping voltage profiles of the new high-rate materials presented herein provide an opportunity for the modelling and electrochemical engineering/industrial communities to develop BMS based on sloping voltage profiles, which may prove to be a substantial commercial advantage for $Nb_{16}W_5O_{55}$, $Nb_{18}W_{16}O_{93}$ and related materials in the growing area of high-power/fast-charging applications.

High-voltage anode materials, such as the niobium tungsten oxides, have several further advantages. Above 1.0 V versus Li^+/Li^0, the formation of solid–electrolyte interphase is minimal, which means that Li will not be lost into side reactions with the electrolyte. Furthermore, in a full cell against, for example, $LiFePO_4$, $LiN(CF_3SO_2)_2$ can be used to replace the more toxic $LiPF_6$ electrolyte salt without degrading Al current collector(s)[89], and Al can be used as the current collector instead of the more expensive Cu while avoiding LiAl alloying potentials (≤0.3 V versus Li^+/Li)[90]. Another incentive to find replacements for nano-$Li_4Ti_5O_{12}$ is the issue of gas evolution and associated swelling or pressure build-up[91,92], which stems from heterogeneous catalysis between the metal-oxide surface and the organic electrolyte. The small particle sizes required to compensate for poor Li^+ and electronic diffusion in $Li_4Ti_5O_{12}$ increase the reactive surface area, and the much smaller surface area of the ~5-μm $Nb_{16}W_5O_{55}$ or $Nb_{18}W_{16}O_{93}$ particles potentially reduces side reactions.

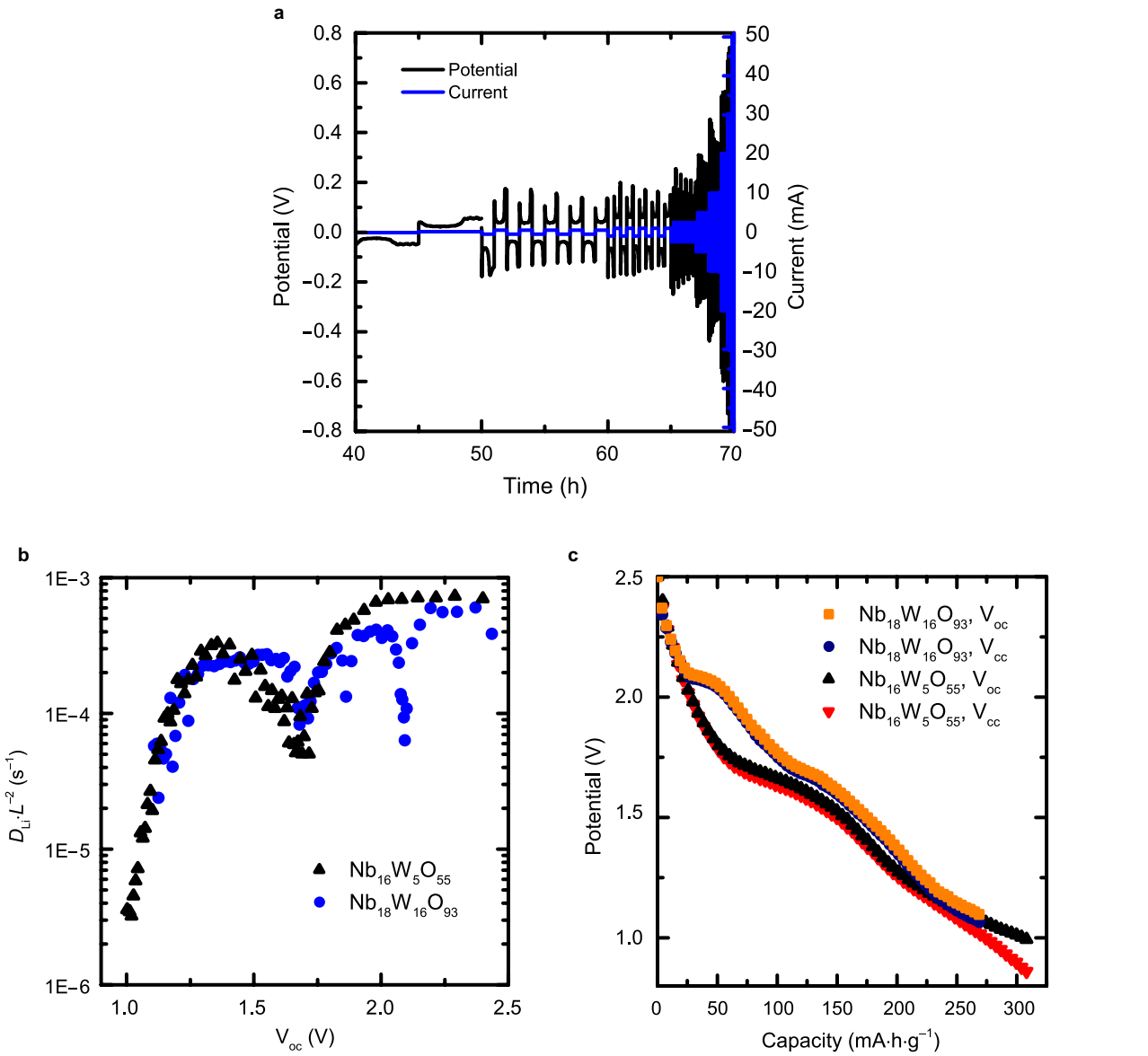
Data availability. The data that support the findings of this study are available from www.repository.cam.ac.uk and the corresponding author upon reasonable request.

55. Roth, R. S. Thermal stability of long range order in oxides. *Prog. Solid State Chem.* **13**, 159–192 (1980).
56. Eyring L. & O'Keefe M. (eds) *The Chemistry of Extended Defects in Non-Metallic Solids* (North-Holland, Amsterdam, 1970).
57. Toby, B. H. & Von Dreele, R. B. GSAS-II: the genesis of a modern open-source all purpose crystallography software package. *J. Appl. Cryst.* **46**, 544–549 (2013).
58. Allpress, J. G. & Roth, R. S. The effect of annealing on the concentration of Wadsley defects in the Nb_2O_5–WO_3 system. *J. Solid State Chem.* **3**, 209–216 (1971).
59. Ragone, D. V. Review of battery systems for electrically powered vehicles. In Proc. Mid-Year Meeting of the Society of Automotive Engineers (SAE, 1968).
60. Rohatgi, A. *WebPlotDigitizer* https://automeris.io/WebPlotDigitizer (2017).
61. Borkiewicz, O. J. et al. The AMPIX electrochemical cell: a versatile apparatus for *in situ* X-ray scattering and spectroscopic measurements. *J. Appl. Cryst.* **45**, 1261–1269 (2012).
62. Caglioti, G., Paoletti, A., Ricci, F. P. Choice of collimators for a crystal spectrometer for neutron diffraction. *Nucl. Instrum.* **3**, 223–228 (1958).
63. Ravel, B. & Newville, M. ATHENA, ARTEMIS, HEPHAESTUS: data analysis for X-ray absorption spectroscopy using IFEFFIT. *J. Synchrotron Radiat.* **12**, 537–541 (2005).

© 2018 Springer Nature Limited. All rights reserved.

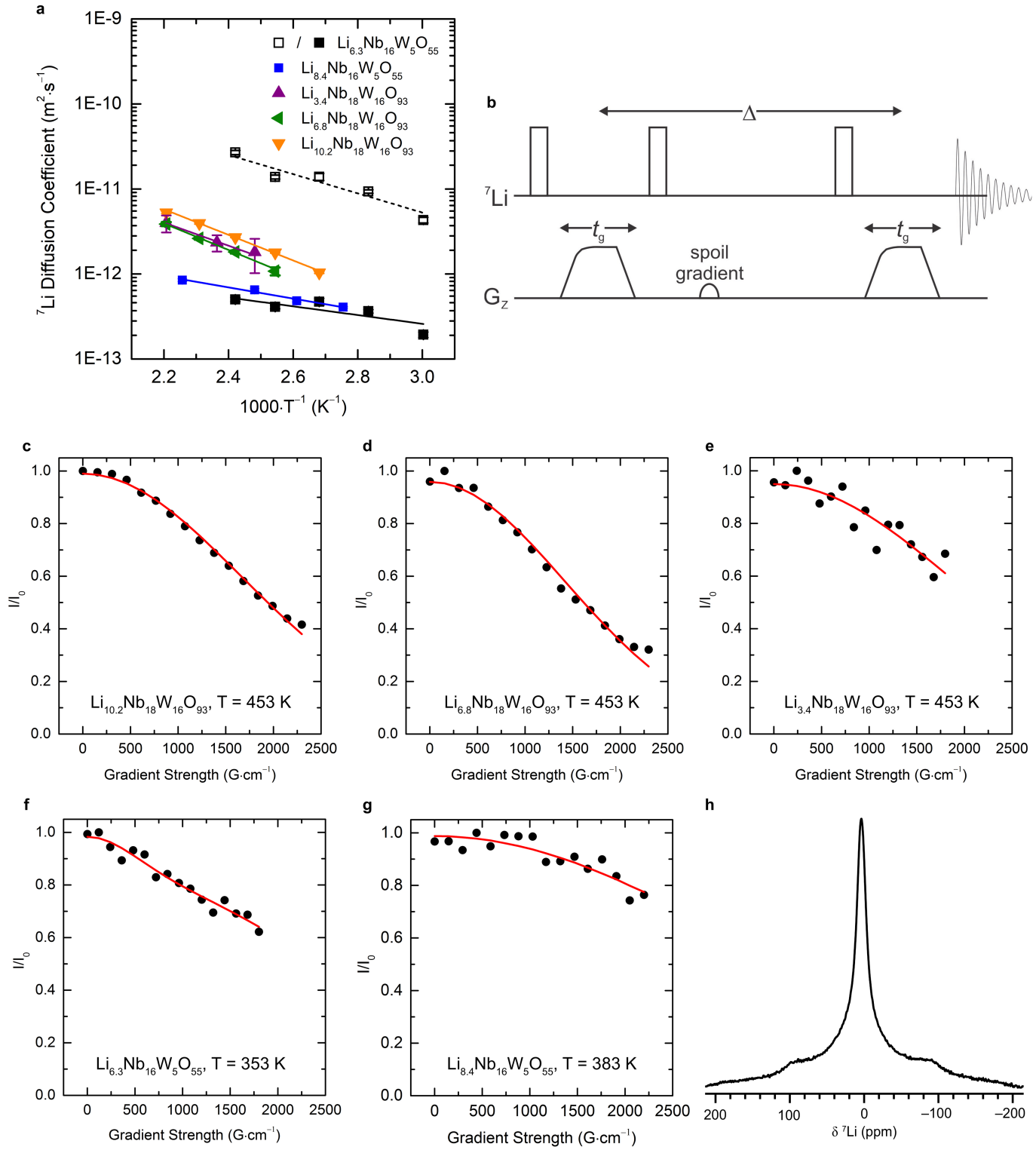
64. Newville, M. IFEFFIT: interactive XAFS analysis and FEFF fitting. *J. Synchrotron Radiat.* **8**, 322–324 (2001).
65. Thompson, A. et al. *X-ray Data Booklet* (Lawrence Berkeley National Laboratory, Berkeley, 2009).
66. Yamamoto, T., Orita, A. & Tanaka, T. Structural analysis of tungsten–zirconium oxide catalyst by W K-edge and L_1-edge XAFS. *X-ray Spectrom.* **37**, 226–231 (2008).
67. Tougerti, A. et al. XANES study of rhenium oxide compounds at the L_1 and L_3 absorption edges. *Phys. Rev. B* **85**, 125136 (2012).
68. Jayarathne, U. et al. X-ray absorption spectroscopy systematics at the tungsten L-edge. *Inorg. Chem.* **53**, 8230–8241 (2014).
69. Drube, W., Treusch, R., Sham, T. K., Bzowski, A. & Soldatov, A. V. Sublifetime-resolution Ag L_3-edge XANES studies of Ag-Au alloys. *Phys. Rev. B* **58**, 6871–6876 (1998).
70. Bersuker, I. B. *The Jahn–Teller Effect* (Cambridge Univ. Press, Cambridge, 2006).
71. Whittle, T. A., Schmid, S. & Howard, C. J. Octahedral tilting in the tungsten bronzes. *Acta Crystallogr. B* **71**, 342–348 (2015).
72. Yan, L. et al. Electrospun $WNb_{12}O_{33}$ nanowires: superior lithium storage capability and their working mechanism. *J. Mater. Chem. A* **5**, 8972–8980 (2017).
73. Stejskal, E. O. & Tanner, J. E. Spin diffusion measurements: spin echoes in the presence of a time-dependent field gradient. *J. Chem. Phys.* **42**, 288–292 (1965).
74. Weppner, W. & Huggins, R. A. Determination of the kinetic parameters of mixed-conducting electrodes and application to the system Li_3Sb. *J. Electrochem. Soc.* **124**, 1569–1578 (1977).
75. Shu, G. J. & Chou, F. C. Sodium-ion diffusion and ordering in single-crystal P2-Na_xCoO_2. *Phys. Rev. B* **78**, 052101 (2008).
76. Reddy, M. V. et al. Studies on the lithium ion diffusion coefficients of electrospun Nb_2O_5 nanostructures using galvanostatic intermittent titration and electrochemical impedance spectroscopy. *Electrochim. Acta* **128**, 198–202 (2014).
77. Ruscher, C., Salje, E. & Hussain, A. The effect of the Nb-W distribution on polaronic transport in ternary Nb-W oxides: electrical and optical properties. *J. Phys. C* **21**, 4465–4480 (1988).
78. Cava, R. J. et al. Electrical and magnetic properties of $Nb_2O_{5-\delta}$ crystallographic shear structures. *Phys. Rev. B* **44**, 6973–6981 (1991).
79. Dickens, P. G. & Whittingham, M. S. The tungsten bronzes and related compounds. *Q. Rev. Chem. Soc.* **22**, 30–44 (1968).
80. Hull, S. Superionics: crystal structures and conduction processes. *Rep. Prog. Phys.* **67**, 1233–1314 (2004).
81. Kamaya, N. et al. A lithium superionic conductor. *Nat. Mater.* **10**, 682–686 (2011).
82. Kuhn, A., Duppel, V. & Lotsch, B. V. Tetragonal $Li_{10}GeP_2S_{12}$ and Li_7GePS_8 – exploring the Li ion dynamics in LGPS Li electrolytes. *Energy Environ. Sci.* **6**, 3548–3552 (2013).
83. Mo, Y., Ong, S. P. & Ceder, G. First principles study of the $Li_{10}GeP_2S_{12}$ lithium super ionic conductor material. *Chem. Mater.* **24**, 15–17 (2012).
84. Phani Dathar, G. K., Balachandran, J., Kent, P. R. C., Rondinone, A. J. & Ganesh, P. Li-ion site disorder driven superionic conductivity in solid electrolytes: a first-principles investigation of β-Li_3PS_4. *J. Mater. Chem. A* **5**, 1153–1159 (2017).
85. Bevan, D. J. M. & Hagenmuller, D. *Non-Stoichiometric Compounds: Tungsten Bronzes, Vanadium Bronzes and Related Compounds 1st edn* (Pergamon Press, Exeter, 1973).
86. Pinus, I., Catti, M., Ruffo, R., Salamone, M. M. & Mari, C. M. Neutron diffraction and electrochemical study of $FeNb_{11}O_{29}$/$Li_{11}FeNb_{11}O_{29}$ for lithium battery anode applications. *Chem. Mater.* **26**, 2203–2209 (2014).
87. Galy, J. & Andersson, S. Structure cristalline de $MoNb_{15}O_{40}F$. *Acta Crystallogr. B* **24**, 1027–1031 (1968).
88. Idrees, F. et al. Facile synthesis of novel Nb_3O_7F nanoflowers, their optical and photocatalytic properties. *CrystEngComm* **15**, 8146–8152 (2013).
89. Zaghib, K., Mauger, A., Groult, H., Goodenough, J. B. & Julien, C. M. Advanced electrodes for high power Li-ion batteries. *Materials* **6**, 1028–1049 (2013).
90. Wen, C. J., Boukamp, B. A., Huggins, R. A. & Weppner, W. Thermodynamic and mass transport properties of "LiAl". *J. Electrochem. Soc.* **126**, 2258–2266 (1979).
91. He, Y.-B. et al. Gassing in $Li_4Ti_5O_{12}$-based batteries and its remedy. *Sci. Rep.* **2**, 913 (2012).
92. Lv, W., Gu, J., Niu, Y., Wen, K. & He, W. Review—gassing mechanism and suppressing solutions in $Li_4Ti_5O_{12}$-based lithium-ion batteries. *J. Electrochem. Soc.* **164**, A2213–A2224 (2017).
93. Vinod Chandran, C. & Heitjans, P. Solid-state NMR studies of lithium ion dynamics across materials classes. *Ann. Rep. NMR Spectrosc.* **89**, 1–102 (2016).
94. Wang, Z. et al. Lithium diffusion in lithium nitride by pulsed-field gradient NMR. *Phys. Chem. Chem. Phys.* **14**, 13535–13538 (2012).
95. Kuhn, A. et al. A new ultrafast superionic Li-conductor: ion dynamics in $Li_{11}Si_2PS_{12}$ and comparison with other tetragonal LGPS-type electrolytes. *Phys. Chem. Chem. Phys.* **16**, 14669–14674 (2014).
96. Kaus, M. et al. Local structures and Li ion dynamics in a $Li_{10}SnP_2S_{12}$-based composite observed by multinuclear solid-state NMR spectroscopy. *J. Phys. Chem. C* **121**, 23370–23376 (2017).
97. Hayamizu, K. & Aihara, Y. Lithium ion diffusion in solid electrolyte $(Li_2S)_7(P_2S_5)_3$ measured by pulsed-gradient spin-echo ^{7}Li NMR spectroscopy. *Solid State Ion.* **238**, 7–14 (2013).
98. Gobet, M., Greenbaum, S., Sahu, G. & Liang, C. Structural evolution and Li dynamics in nanophase Li_3PS_4 by solid-state and pulsed-field gradient NMR. *Chem. Mater.* **26**, 3558–3564 (2014).
99. Hayamizu, K. et al. NMR studies on lithium ion migration in sulfide-based conductors, amorphous and crystalline Li_3PS_4. *Solid State Ion.* **285**, 51–58 (2016).
100. Holzmann, T. et al. $Li_{0.6}[Li_{0.2}Sn_{0.8}S_2]$ – a layered lithium superionic conductor. *Energy Environ. Sci.* **9**, 2578–2585 (2016).
101. Ishiyama, H. et al. Nanoscale diffusion tracing by radioactive ^{8}Li tracer. *Jpn. J. Appl. Phys.* **53**, 110303 (2014).
102. Holz, M. & Weingartner, H. Calibration in accurate spin-echo self-diffusion measurements using ^{1}H and less-common nuclei. *J. Magn. Reson.* **92**, 115–125 (1991).
103. Hayamizu, K. Temperature dependence of self-diffusion coefficients of ions and solvents in ethylene carbonate, propylene carbonate, and diethyl carbonate single solutions and ethylene carbonate + diethyl carbonate binary solutions of $LiPF_6$ studied by NMR. *J. Chem. Eng. Data* **57**, 2012–2017 (2012).
104. Chowdhury, M. T., Takekawa, R., Iwai, Y., Kuwata, N. & Kawamura, J. Lithium ion diffusion in Li β-alumina single crystals measured by pulsed field gradient NMR spectroscopy. *J. Chem. Phys.* **140**, 124509 (2014).
105. Hayamizu, K. & Seki, S. Long-range Li ion diffusion in NASICON-type $Li_{1.5}Al_{0.5}Ge_{1.5}(PO_4)_3$ (LAGP) studied by ^{7}Li pulsed-gradient spin-echo NMR. *Phys. Chem. Chem. Phys.* **19**, 23483–23491 (2017).
106. Arbi, K. et al. Ionic mobility in Nasicon-type $LiM^{IV}2(PO_4)_3$ materials followed by ^{7}Li NMR spectroscopy. *MRS Proc.* **1313** (2011).
107. Hayamizu, K., Matsuda, Y., Matsui, M. & Imanishi, N. Lithium ion diffusion measurements on a garnet-type solid conductor $Li_{6.6}La_3Zr_{1.6}Ta_{0.4}O_{12}$ by using a pulsed-gradient spin-echo NMR method. *Solid State Nucl. Magn. Reson.* **70**, 21–27 (2015).
108. Kuhn, A. et al. Li self-diffusion in garnet-type $Li_7La_3Zr_2O_{12}$ as probed directly by diffusion-induced ^{7}Li spin-lattice relaxation NMR spectroscopy. *Phys. Rev. B* **83**, 094302 (2011).
109. Langer, J., Epp, V., Heitjans, P., Mautner, F. A. & Wilkening, M. Lithium motion in the anode material LiC_6 as seen via time-domain ^{7}Li NMR. *Phys. Rev. B* **88**, 094304 (2013).
110. Mali, M., Roos, J., Sonderegger, M., Brinkmann, D. & Heitjans, P. ^{6}Li and ^{7}Li diffusion coefficients in solid lithium measured by the NMR pulsed field gradient technique. *J. Phys. F Met. Phys.* **18**, 403–412 (1988).
111. Sugiyama, J. et al. Li-ion diffusion in $Li_4Ti_5O_{12}$ and $LiTi_2O_4$ battery materials detected by muon spin spectroscopy. *Phys. Rev. B* **92**, 014417 (2015).
112. Sugiyama, J. et al. Lithium diffusion in spinel $Li_4Ti_5O_{12}$ and $LiTi_2O_4$ films detected with ^{8}Li beta-NMR. *Phys. Rev. B* **96**, 094402 (2017).
113. Wilkening, M. et al. Microscopic Li self-diffusion parameters in the lithiated anode material $Li_{4+x}Ti_5O_{12}$ ($0 \leq x \leq 3$) measured by ^{7}Li solid state NMR. *Phys. Chem. Chem. Phys.* **9**, 6199–6202 (2007).
114. Ruprecht, B., Wilkening, M., Uecker, R. & Heitjans, P. Extremely slow Li ion dynamics in monoclinic Li_2TiO_3—probing macroscopic jump diffusion via ^{7}Li NMR stimulated echoes. *Phys. Chem. Chem. Phys.* **14**, 11974–11980 (2012).
115. Wagemaker, M., van de Krol, R., Kentgens, A. P. M., van Well, A. A. & Mulder, F. M. Two phase morphology limits lithium diffusion in TiO_2 (anatase): a ^{7}Li MAS NMR study. *J. Am. Chem. Soc.* **123**, 11454–11461 (2001).
116. Wagemaker, M. et al. The influence of size on phase morphology and Li-ion mobility in nanosized lithiated anatase TiO_2. *Chem. Eur. J.* **13**, 2023–2028 (2007).
117. Verhoeven, V. W. J. et al. Lithium dynamics in $LiMn_2O_4$ probed directly by two-dimensional ^{7}Li NMR. *Phys. Rev. Lett.* **86**, 4314–4317 (2001).
118. Ishiyama, H. et al. Direct measurement of nanoscale lithium diffusion in solid battery materials using radioactive tracer of ^{8}Li. *Nucl. Instrum. Methods B* **376**, 379–381 (2016).
119. Bork, D. & Heitjans, P. NMR relaxation study of ion dynamics in nanocrystalline and polycrystalline $LiNbO_3$. *J. Phys. Chem. B* **102**, 7303–7306 (1998).
120. Ruprecht, B. & Heitjans, P. Ultraslow lithium diffusion in Li_3NbO_4 probed by ^{7}Li stimulated echo NMR spectroscopy. *Diffusion Fundamentals* **12**, 100–101 (2010).
121. Sale, M. & Avdeev, M. 3DBVSMAPPER: a program for automatically generating bond-valence sum landscapes. *J. Appl. Crystallogr.* **45**, 1054–1056 (2012).
122. Avdeev, M., Sale, M., Adams, S. & Rao, R. P. Screening of the alkali-metal ion containing materials from the Inorganic Crystal Structure Database (ICSD) for high ionic conductivity pathways using the bond valence method. *Solid State Ion.* **225**, 43–46 (2012).
123. Brown, I. D. *The Chemical Bond in Inorganic Chemistry: The Bond Valence Model* (Oxford Univ. Press, Oxford, 2002).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 1 | Overpotential in Li||Li symmetric cells and GITT of niobium tungsten oxides. a, Li||Li symmetric cells were configured identically to those used for metal-oxide testing, with the exception of a second Li disk replacing the composite electrode. Rate testing was carried out with current densities corresponding to the rates shown in Fig. 2a–e, with 5 cycles at 100 μA (C/5), 500 μA (1C), 1 mA (2C), 2.5 mA (5C) and 5 mA (10C) and 10 cycles at 10 mA (20C), 20 mA (40C), 30 mA (60C) and 50 mA (100C). The 'rates' in parentheses indicate the inverse of the time (in hours) that the current was applied, simulating the conditions (current densities, periods of applied current and thus total charge transferred) of the rate test experiments. An excerpt of the results is shown here from the fifth cycle at C/5 until the end of the test; the test was performed twice with the same overpotentials observed in both cells. At low current densities, below 1 mA (2C), the overpotential is below 100 mV; however, at 5 mA (10C) the overpotential rises to 200 mV and increases to about 700 mV at 100C. **b**, **c**, Relative changes in lithium diffusion as a function of open-circuit voltage (V_{oc}; **b**) and open-circuit voltage versus closed-circuit voltage (V_{cc}; **c**) from the GITT measurements. The plot in **c** shows the 'thermodynamic' electrochemical profiles at C/20 with a 12-h rest period at each point, reaching a full discharge in approximately one month. In $Nb_{16}W_5O_{55}$, the fastest diffusion is observed from the dilute limit to $Li_{4.5(5)}Nb_{16}W_5O_{55}$, dropping by two orders of magnitude in the low-voltage window, where more than 1 Li^+/TM are incorporated. The GITT data indicate that the second electrochemical region of $Nb_{16}W_5O_{55}$ is broader than typically observed in a two-phase reaction, but the observed discontinuity in the $D_{Li}L^{-2}$ values in this region suggests that $Nb_{16}W_5O_{55}$ approaches two-phase behaviour. The average diffusion coefficient in $Nb_{18}W_{16}O_{93}$ is similar to that of $Nb_{16}W_5O_{55}$. The bronze also displays discontinuities in $D_{Li}L^{-2}$ at 2.1, 1.85 and 1.7 V. In both phases, the low-voltage region (below 1.25 V, well over 1 Li^+/TM) is characterized by an increasing overpotential and suppressed kinetics.

© 2018 Springer Nature Limited. All rights reserved.

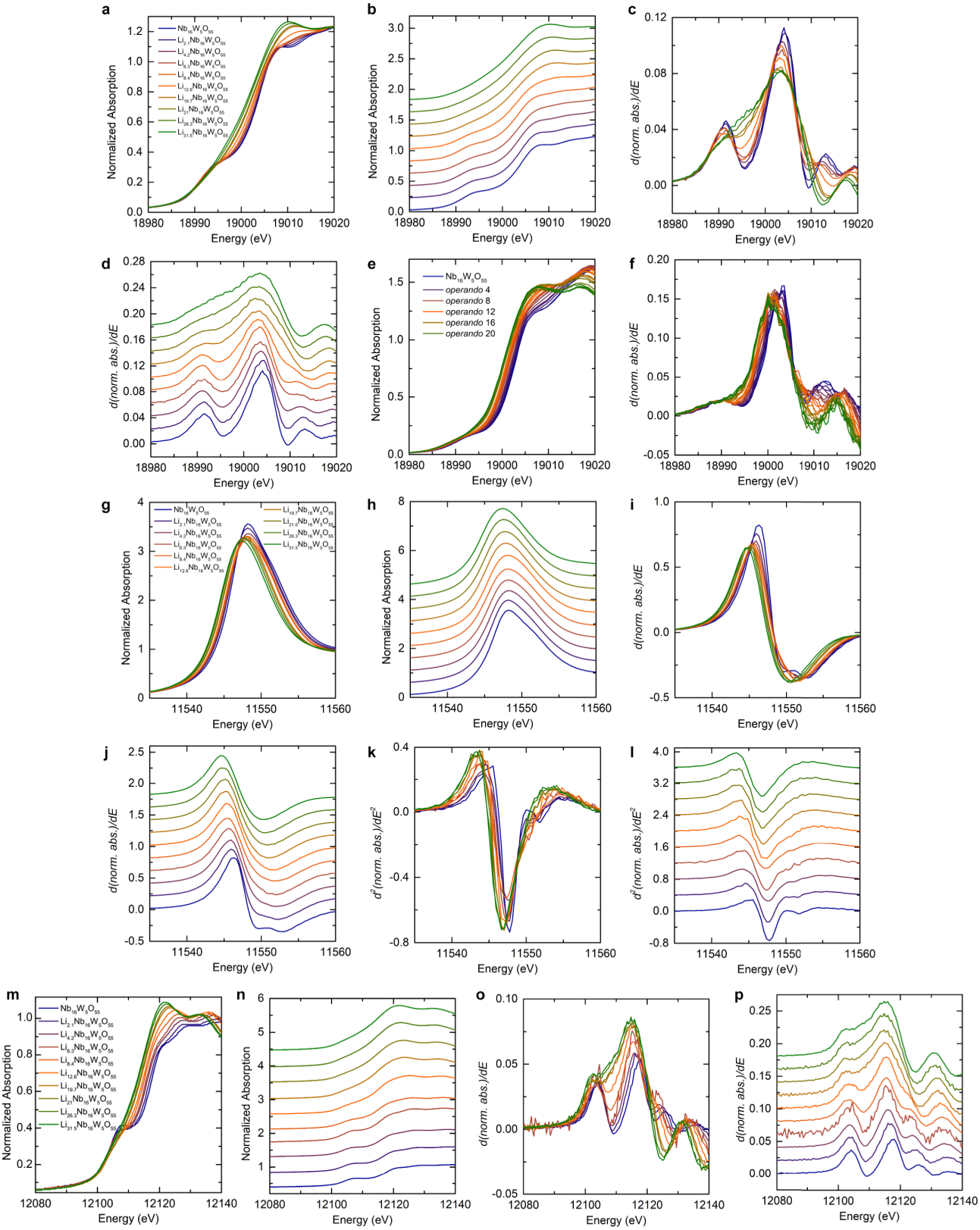


Extended Data Fig. 2 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

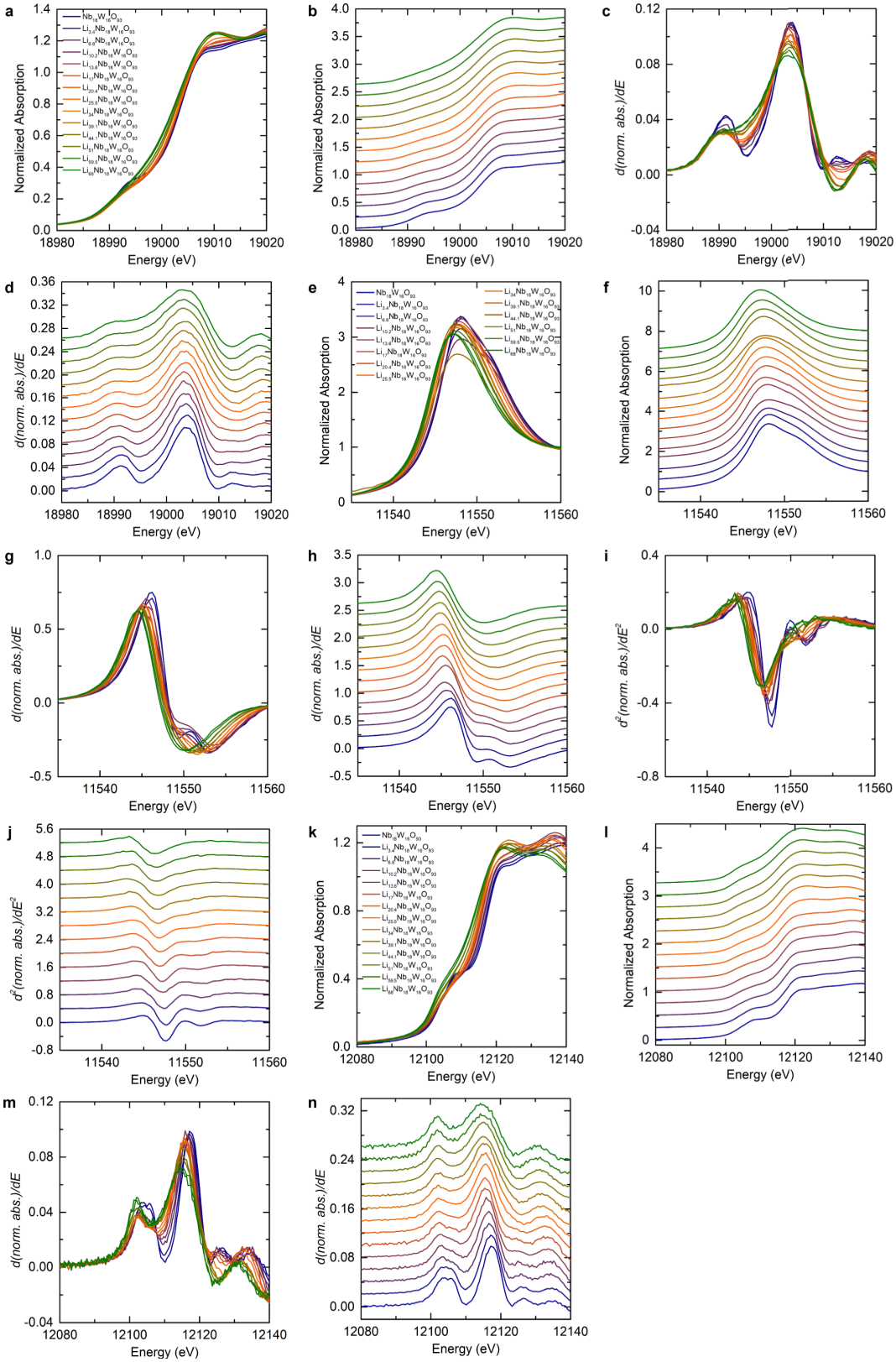
Extended Data Fig. 2 | Lithium diffusion from ^{7}Li PFG NMR spectroscopy. a, The lithium diffusion coefficients of $Li_xNb_{16}W_5O_{55}$ ($x = 6.3$, 8.4) and $Li_xNb_{18}W_{16}O_{93}$ ($x = 3.4$, 6.8, 10.2) were measured in the temperature range 333–453 K (Table 1). The filled (85% signal contribution) and empty (15% signal contribution) symbols for $Li_{6.3}Nb_{16}W_5O_{55}$ correspond to the observed two-component diffusion (see main text and Methods). In most cases, the error bars (2σ, obtained from the fit) are smaller than the sizes of the points. **b**, Stimulated-echo PFG sequence used to measure ^{7}Li diffusivities, showing both radiofrequency (^{7}Li) and magnetic-field-gradient (G_z) pulses. Here, the gradient pulse duration (t_g) includes the up-ramping, time on and down-ramping of the opt composite gradient pulses. During Δ, a short spoiler gradient was used to remove residual transverse magnetization. **c–g**, Representative ^{7}Li decay curves showing the normalized NMR signal intensity as a function of gradient strength for the bronze structures $Li_{10.2}Nb_{18}W_{16}O_{93}$ at 453 K (**c**), $Li_{6.8}Nb_{18}W_{16}O_{93}$ at 453 K (**d**) and $Li_{3.4}Nb_{18}W_{16}O_{93}$ at 453 K (**e**) and the block structures $Li_{6.3}Nb_{16}W_5O_{55}$ at 353 K (**f**) and $Li_{8.4}Nb_{16}W_5O_{55}$ at 383 K (**g**). Black circles represent experimental data points and red lines represent mono- (**c–e**, **g**) or biexponential (**f**) fits to the data by the Stejskal–Tanner equation. Biexponential fits to all samples except $Li_{6.3}Nb_{16}W_5O_{55}$ did not lead to improved fits over those obtained with monoexponential fits. The poor signal-to-noise ratio of $Li_{6.3}Nb_{16}W_5O_{55}$, even with more than 100 mg of electrode sample, did not allow us to explore alternative models beyond the mono- and biexponential models, such as models considering the effect of anisotropic diffusion, which manifests as subtle differences in echo attenuation[94]. **h**, A representative one-dimensional ^{7}Li NMR spectrum ($Li_{8.4}Nb_{16}W_5O_{55}$, static, 403 K) showing the shift (δ ^{7}Li) region and lineshape.

© 2018 Springer Nature Limited. All rights reserved.



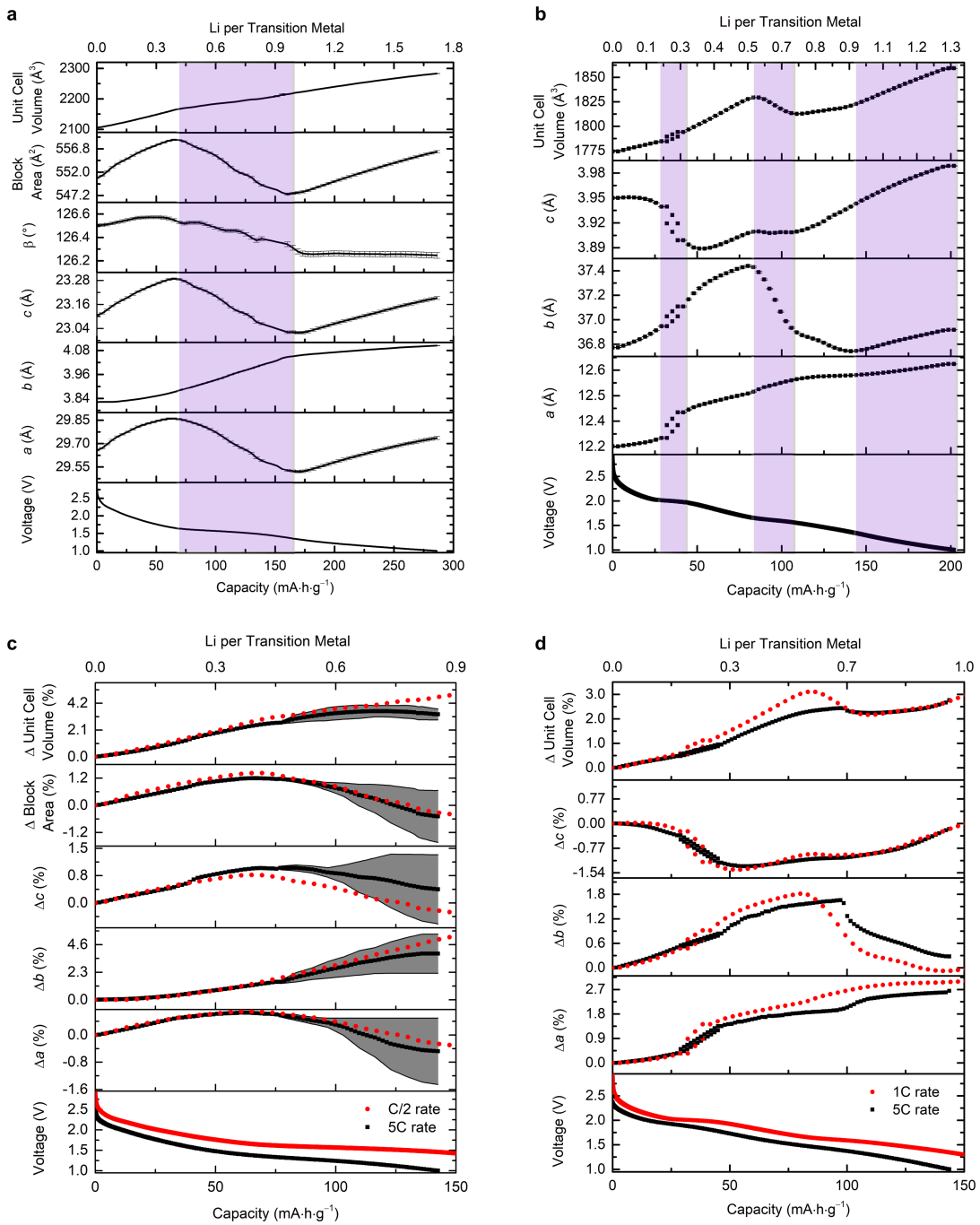
Extended Data Fig. 3 | $Nb_{16}W_5O_{55}$ X-ray absorption spectroscopy. **a**–**d**, $Li_xNb_{16}W_5O_{55}$ ex situ (see Supplementary Fig. 14) Nb K edge XANES spectra (**a**, **b**) and derivative spectra (**c**, **d**). **e**, **f**, Operando Nb K edge XANES spectra of $Nb_{16}W_5O_{55}$ and 22 discharge spectra, with each successive spectrum at about +11 mA h g^{-1} (**e**), and operando derivative spectra (**f**; 23 spectra shown, 5 colours labelled). The pre-edge and main edge are at about 18,991 and 19,004 eV, respectively. Spectra in **b** and **d** are vertically offset by 0.2 and 0.02, respectively, for clarity. **g**–**l**, $Li_xNb_{16}W_5O_{55}$ ex situ W L_{II} edge XANES spectra (**g**, **h**), derivative spectra (**i**, **j**) and second-derivative spectra (**k**, **l**). Spectra in **h**, **j** and **l** are vertically offset by 0.5, 0.2 and 0.4, respectively, for clarity. **m**–**p**, $Li_xNb_{16}W_5O_{55}$ ex situ W L_I edge XANES spectra (**m**, **n**) and derivative spectra (**o**, **p**) near the W L_I absorption edge. The pre-edge is at about 12,104 eV and the main edge at about 12,117 eV. Spectra are vertically offset by 0.5 in **n** and by 0.02 in **p** for clarity.

© 2018 Springer Nature Limited. All rights reserved.



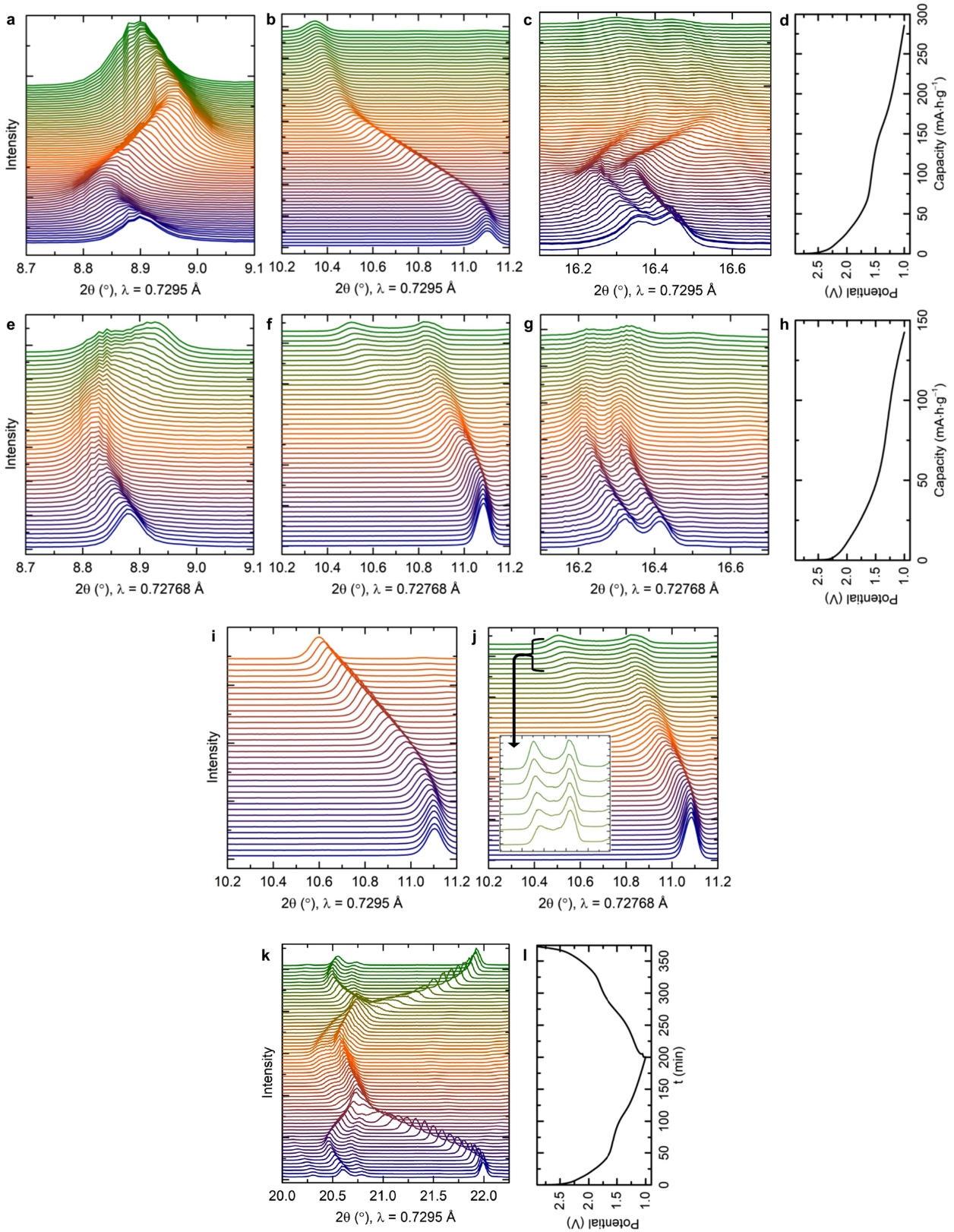
Extended Data Fig. 4 | $Nb_{18}W_{16}O_{93}$ X-ray absorption spectroscopy. **a–d**, $Li_xNb_{18}W_{16}O_{93}$ ex situ Nb K edge XANES spectra (**a**, **b**) and derivative spectra (**c, d**). The pre-edge and main edge are at about 18,991 and 19,004 eV, respectively. Spectra in **b** and **d** are vertically offset by 0.2 and 0.02, respectively, for clarity. **e–j**, $Li_xNb_{18}W_{16}O_{93}$ ex situ W L_{II} edge XANES spectra (**e**, **f**), derivative spectra (**g, h**) and second-derivative spectra (**i, j**). Spectra in **f**, **h** and **j** are vertically offset by 0.5, 0.2 and 0.4, respectively, for clarity. **k–m**, $Li_xNb_{18}W_{16}O_{93}$ ex situ W L_I edge XANES spectra (**k**, **l**) and derivative spectra (**m, n**) near the W L_I absorption edge. The pre-edge is at about 12,104 eV and the main edge at about 12,117 eV. Spectra are vertically offset by 0.25 in **l** and by 0.02 in **n** for clarity.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 5 | Lattice evolution of $Nb_{16}W_5O_{55}$ and $Nb_{18}W_{16}O_{93}$ upon lithiation. a, **b**, Absolute lattice parameter values resulting from Rietveld refinement of operando diffraction data, analogous to that of Fig. 4c, $Nb_{16}W_5O_{55}$ (**a**) and Fig. 4d, $Nb_{18}W_{16}O_{93}$ (**b**). Error bars show the standard deviation of each parameter, as estimated from the fits (approximately equal to the symbol height in **b**). Shading distinguishes the different structural regions. For $Nb_{18}W_{16}O_{93}$, the second stage (two-phase region) contains two sets of lattice parameters. **c**, Structure evolution of $Nb_{16}W_5O_{55}$ as a function of rate. At high rates, the lithiation reaction becomes inhomogeneous (Extended Data Fig. 6) and cannot be fitted with a single set of lattice parameters over the whole electrode. The shaded grey area for the 5C data corresponds to the range of each unit-cell parameter. **d**, Structure evolution of $Nb_{18}W_{16}O_{93}$ as a function of rate. The mechanism of $Li_xNb_{18}W_{16}O_{93}$ lattice evolution does not appear to be strongly rate-dependent. The reaction extends further at lower rates within the same voltage range, partly because of a smaller overpotential at lower current densities (Extended Data Fig. 1a).

© 2018 Springer Nature Limited. All rights reserved.

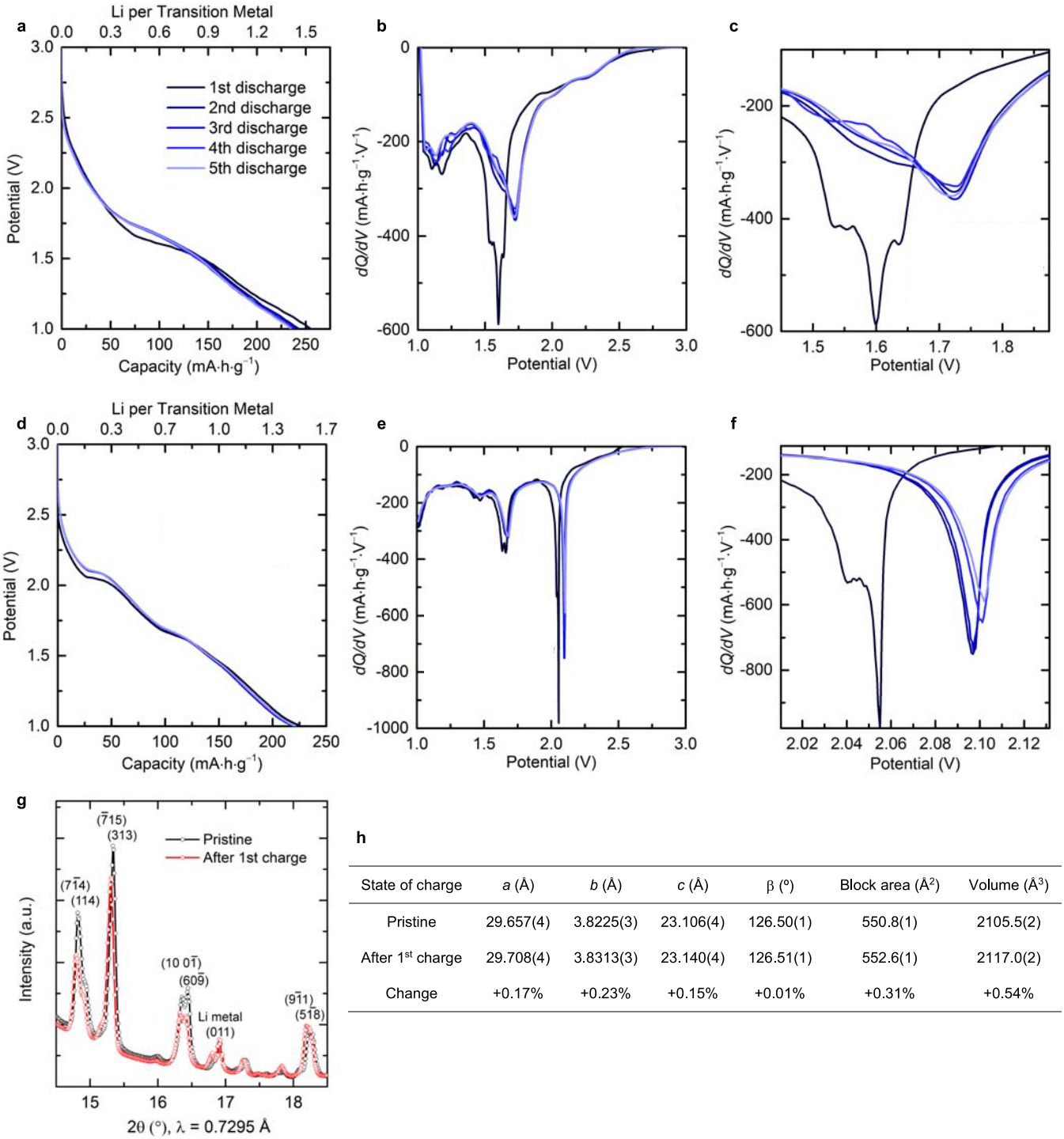


Extended Data Fig. 6 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 6 | Solid-solution structure evolution and reversibility of $Nb_{16}W_5O_{55}$. **a**–**c**, Profile evolution of selected reflections at C/2 and **d**, the corresponding electrochemical discharge profile. **e**–**g**, Profile evolution of selected reflections at 5C and **h**, the corresponding electrochemical discharge profile. **a**, **e**, $(60\overline{4})$. **b**, **f**, overlapping (110) and $(11\bar{1})$. **c**, **g**, $(10\ 0\bar{1})$ (left) and $(60\overline{9})$ (right), with smaller overlapping reflections. The evolution in each case commences analogously; as lithiation increases, the structure evolution at high rate becomes inhomogeneous and the contraction of the *ac* block is not fully realized. **i**, The overlapping (110) and $(11\bar{1})$ reflections shift smoothly to larger *d* spacing (smaller 2θ) for the entire sample at C/2, whereas in **j**, there is substantial intensity across a range of 2θ values at high rate. This inter-peak intensity is shown more clearly in the inset of **j**. The intensity range represents a range of lattice parameters (Extended Data Fig. 5c) and an inhomogeneous solid-solution reaction probably resulting from inhomogeneous lithium transport and concentration gradients within the electrolyte[33,34]. Diffraction patterns are shown from 0 to about 145 mA h g^{-1} (ascending), which corresponds to the full 5C discharge and partial C/2 discharge. **k**, **l**, Structure reversibility over a full electrochemical lithiation/delithiation cycle of $Nb_{16}W_5O_{55}$ at C/2. In **k**, the $(14\ 0\overline{9})$/(407) (left) and (020) (right) reflections are displayed over a full operando synchrotron X-ray diffraction discharge–charge cycle from 3.0 to 1.0 V with multi-redox (de)lithiation. The symmetry of discharge and charge is apparent, although a small amount of lithium remains in the structure after charging (Extended Data Fig. 7).

© 2018 Springer Nature Limited. All rights reserved.

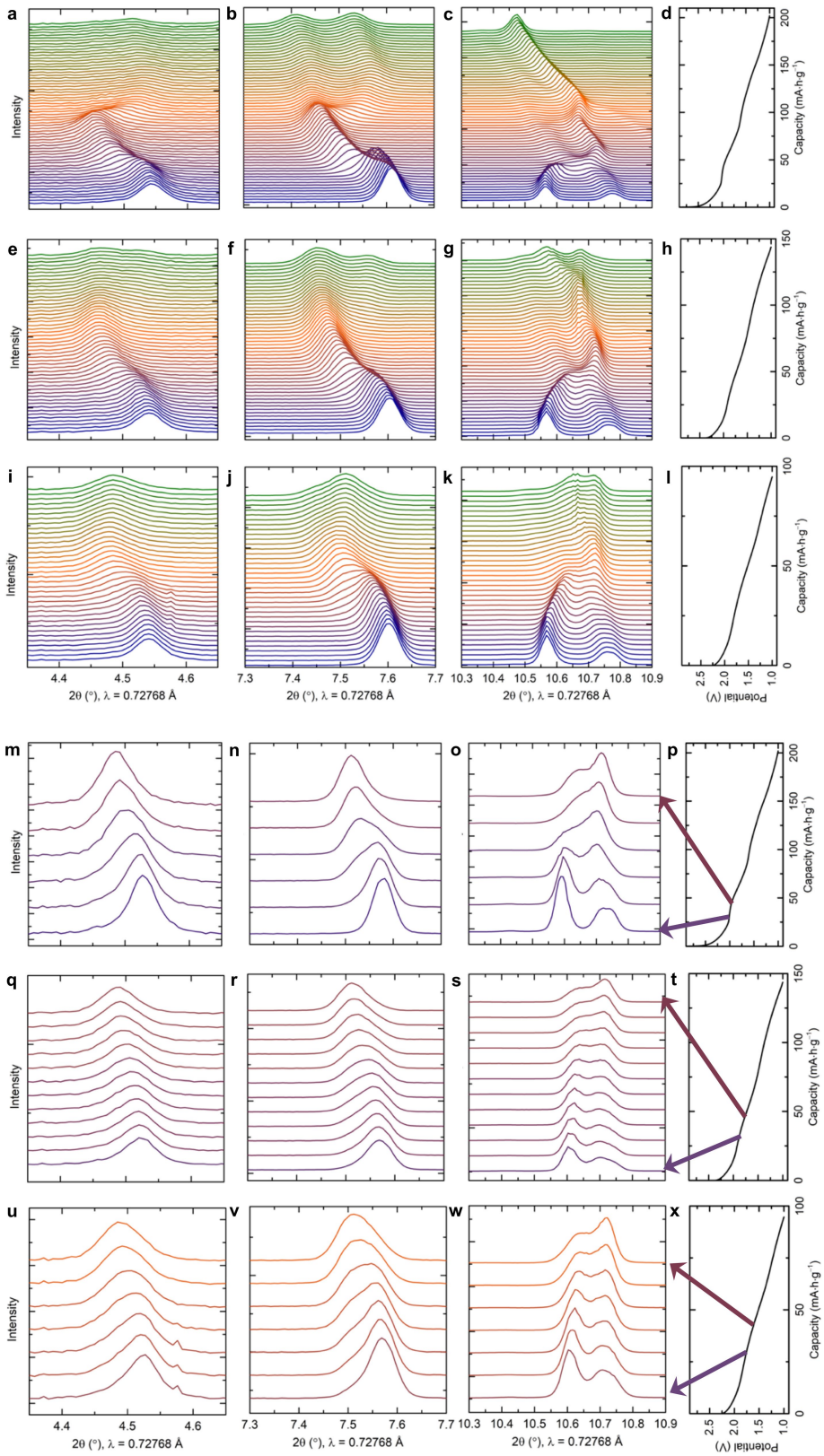


h

State of charge	*a* (Å)	*b* (Å)	*c* (Å)	β (°)	Block area (Å^2)	Volume (Å^3)
Pristine	29.657(4)	3.8225(3)	23.106(4)	126.50(1)	550.8(1)	2105.5(2)
After 1st charge	29.708(4)	3.8313(3)	23.140(4)	126.51(1)	552.6(1)	2117.0(2)
Change	+0.17%	+0.23%	+0.15%	+0.01%	+0.31%	+0.54%

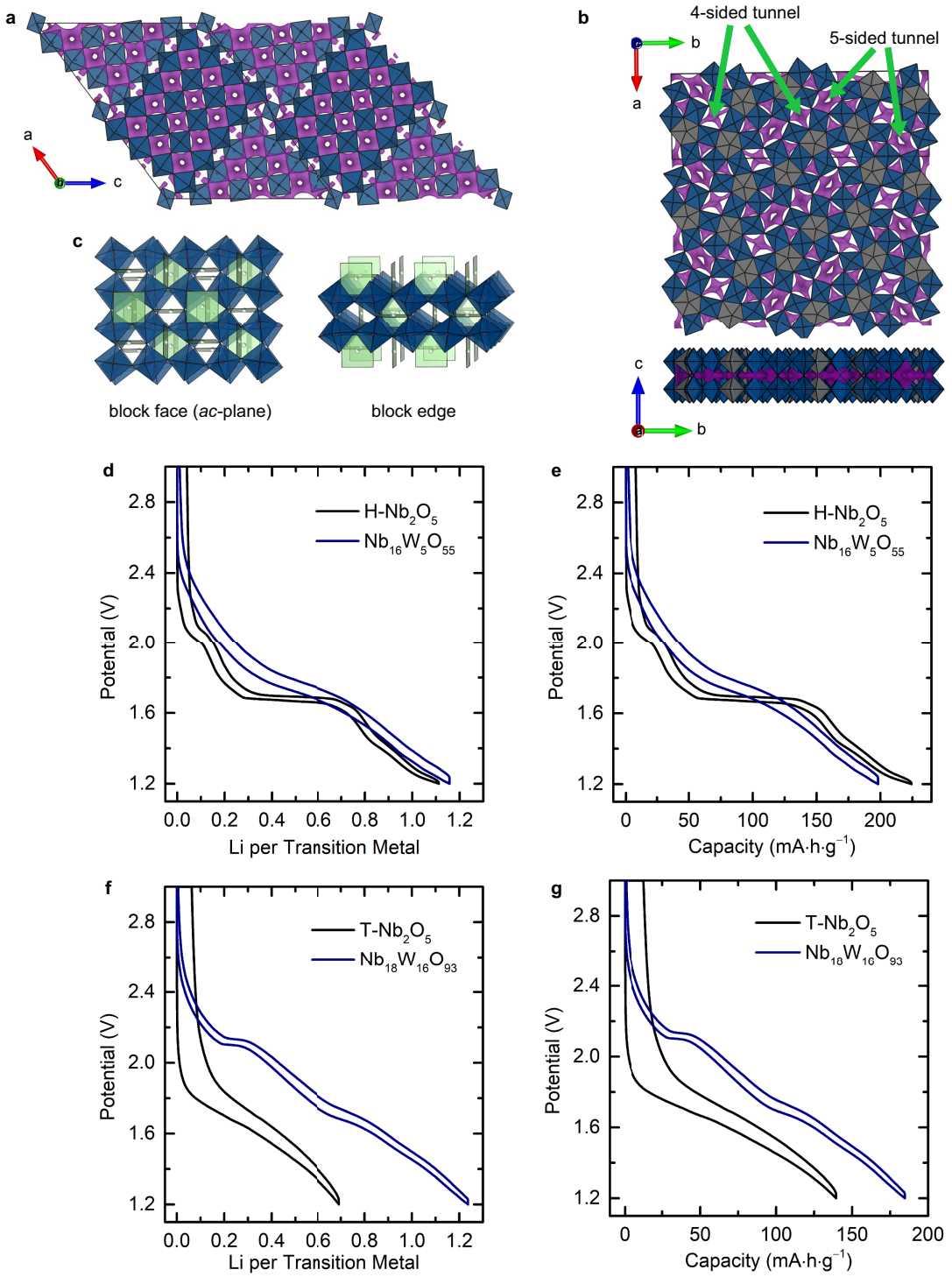
Extended Data Fig. 7 | Electrochemical profile evolution and early-cycle lithium retention of $Nb_{16}W_5O_{55}$ and $Nb_{18}W_{16}O_{93}$. a–f, There is an activation process on the first cycle for $Nb_{16}W_5O_{55}$ (**a–c**) and $Nb_{18}W_{16}O_{93}$ (**d–f**), which extends to a much smaller extent in the next several cycles, leading to an increase in intercalation voltage at the first 'plateau-like' feature and a broadening of the d*Q*/d*V* peaks (**b**, **c**, **e, f**). The phenomenon is associated with a retention of lithium in the structure. The occurrence of this activation phenomenon is structure-independent, indicating that it may have an electronic origin. The electrochemical data in **a–f** were collected at C/5 but the phenomenon is also observed at other rates. **g**, Diffraction patterns of $Nb_{16}W_5O_{55}$ before lithiation and after the first charge from operando measurements at C/2. **h**, Lattice parameters of the pristine and charged structure show changes that indicate that some lithium was retained in the structure after charging the electrode, commensurate with the changes from first- to second-cycle electrochemistry.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 8 | Operando X-ray diffraction patterns of $Nb_{18}W_{16}O_{93}$ from 1C to 10C. a–l, The profile evolution of selected reflections and the corresponding electrochemical discharge profile is shown at 1C (**a**–**d**), 5C (**e**–**h**) and 10C (**i**–**l**). **a**, **e**, **i**, (040); **b**, **f**, **j**, overlapped (230) and (160); **c**, **g**, **k**, (001) (left, initially) and overlapped (330) and (190) (right, initially). The evolution in each case is similar in mechanism and differs in the extent of reaction, consistent with the electrochemical profiles. **m**–**x**, The two-phase region of $Li_xNb_{18}W_{16}O_{93}$ for $x \approx 6.6$–10.2 as a function of rate. Selected regions of the diffraction pattern at shown at 1C (**m**–**p**), 5C (**q**–**t**) and 10C (**u**–**x**). **m**, **q**, **u**, (040); **n**, **r**, **v**, overlapped (230) and (160); **o**, **s**, **w**, (001) (left, initially) and overlapped (330) and (190) (right, initially). The arrows from the electrochemical discharge profiles in **p**, **t** and **x** correspond to the first and last diffraction patterns collected in the two-phase region (about 29–45 mA h g^{-1}), as indicated. Metastable intermediates that can occur at high rates in two-phase systems such as $LiFePO_4$ would be difficult to distinguish in this system owing to the small compositional and structural changes that are associated with the $Li_{6.6}Nb_{18}W_{16}O_{93}$ to $Li_{10.2}Nb_{18}W_{16}O_{93}$ two-phase reaction.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | Prospective lithium positions and pathways in block-type and bronze-type ternary niobium tungsten oxides and electrochemical comparisons to binary niobium oxides. a, **b**, Bond valence sum maps of $Nb_{16}W_5O_{55}$ (**a**) and $Nb_{18}W_{16}O_{93}$ (**b**) depict stable lithium positions and pathways according to bond valence energy landscape calculations performed in 3DBVSMAPPER[121]. Calculations were performed over a fine grid with $149 \times 20 \times 116$ points computed for $Nb_{16}W_5O_{55}$ and $61 \times 184 \times 20$ points computed for $Nb_{18}W_{16}O_{93}$ along their respective crystallographic axes. Isosurface levels are shown at '2.0 eV', which is a parameter used to visualize ionic pathways and not a quantitative estimation. The bond valence sum and bond valence energy landscape provide an indication of lithium positions and diffusion pathways in complex or novel systems and have shown good agreement with experimental and computational investigations of structure and dynamics[121–123]. **c**, Possible intrablock lithium positions for $Nb_{16}W_5O_{55}$ based on Li_xMO_3 in the low-lithium regime, before it undergoes intercalant-induced distortion. **d**, **e**, Block phases $Nb_{16}W_5O_{55}$ and H-Nb_2O_5 are compared on the basis of Li^+/TM (**d**) and gravimetric capacity (**e**) on the third cycle at C/5. **f**, **g**, Bronze-like phases $Nb_{18}W_{16}O_{93}$ and T-Nb_2O_5 are compared on the basis of Li^+/TM (**f**) and gravimetric capacity (**g**) on the third cycle at C/5.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Lithium self-diffusion coefficients of lithium-ion battery electrode materials, solid electrolytes, liquid electrolytes and reference compounds

Compound	Structure Type	D_{Li} ($m^2 \cdot s^{-1}$)	T (K)	Technique	Reference
$Li_{6.3}Nb_{16}W_5O_{55}$	Block, *cs*	2.1×10^{-12} / 1.7×10^{-13}	298	PFG NMR (extrapolated)	This work
$Li_{6.3}Nb_{16}W_5O_{55}$	Block, *cs*	2.7×10^{-11} / 5.2×10^{-13}	413	PFG NMR	This work
$Li_{8.4}Nb_{16}W_5O_{55}$	Block, *cs*	1.6×10^{-13}	298	PFG NMR (extrapolated)	This work
$Li_{8.4}Nb_{16}W_5O_{55}$	Block, *cs*	6.5×10^{-13}	403	PFG NMR	This work
$Li_{3.4}Nb_{18}W_{16}O_{93}$	Bronze-like	1.1×10^{-13}	298	PFG NMR (extrapolated)	This work
$Li_{3.4}Nb_{18}W_{16}O_{93}$	Bronze-like	2.0×10^{-12}	413	PFG NMR	This work
$Li_{6.8}Nb_{18}W_{16}O_{93}$	Bronze-like	1.1×10^{-13}	298	PFG NMR (extrapolated)	This work
$Li_{6.8}Nb_{18}W_{16}O_{93}$	Bronze-like	1.8×10^{-12}	413	PFG NMR	This work
$Li_{10.2}Nb_{18}W_{16}O_{93}$	Bronze-like	1.1×10^{-13}	298	PFG NMR (extrapolated)	This work
$Li_{10.2}Nb_{18}W_{16}O_{93}$	Bronze-like	2.7×10^{-12}	413	PFG NMR	This work
$Li_{10}GeP_2S_{12}$	Thio-LISICON	2×10^{-12}	298	PFG NMR	82
$Li_{10}GeP_2S_{12}$	Thio-LISICON	4×10^{-11}	453	PFG NMR	82
Li_7GePS_8	Thio-LISICON	2×10^{-12}	298	PFG NMR	82
Li_7GePS_8	Thio-LISICON	4×10^{-11}	453	PFG NMR	82
$Li_{10}SnP_2S_{12}$	Thio-LISICON	1.4×10^{-12}	298	PFG NMR	95,96
$Li_{10}SnP_2S_{12}$	Thio-LISICON	4×10^{-11}	453	PFG NMR	95
$Li_{11}Si_2PS_{12}$	Thio-LISICON	3.5×10^{-12}	298	PFG NMR	95
$Li_{11}Si_2PS_{12}$	Thio-LISICON	4×10^{-11}	453	PFG NMR	95
$Li_7P_3S_{11}$	Thio-LISICON	$1–5\times10^{-12}$	303	PFG NMR	97
$Li_7P_3S_{11}$	Thio-LISICON	$2–11\times10^{-12}$	353	PFG NMR	97
β-Li_3PS_4	Thio-LISICON	5.4×10^{-13}	373	PFG NMR	98
amorphous-Li_3PS_4	Amorphous (Thio-LISICON)	6.5×10^{-13}	303	PFG NMR	99
amorphous-Li_3PS_4	Amorphous (Thio-LISICON)	$1.6–3.4\times10^{-12}$	353	PFG NMR	99
$Li_{0.6}[Li_{0.2}Sn_{0.8}S_2]$	Layered (O1)	$2–20\times10^{-12}$	298	PFG NMR	100
$Li_{0.6}[Li_{0.2}Sn_{0.8}S_2]$	Layered (O1)	2.6×10^{-10}	407	PFG NMR	100
$Li_{3.4}V_{0.6}Si_{0.4}O_4$	LISICON	5×10^{-15}	333	tracer	101
0.25 M LiCl (aq.)	Solution	9.6×10^{-10}	298	PFG NMR	102
1.0 M $LiPF_6$ in PC	Liquid electrolyte	8.3×10^{-11}	298	PFG NMR	103
1.0 M $LiPF_6$ in EC	Liquid electrolyte	9.9×10^{-11}	298	PFG NMR	103
1.0 M $LiPF_6$ in DEC	Liquid electrolyte	2.5×10^{-10}	298	PFG NMR	103
Li β-alumina (*ab* plane)	β-alumina	1.2×10^{-11}	298	PFG NMR	104
Li β-alumina (*c*-axis)	β-alumina	6.4×10^{-13}	333	PFG NMR	104
Li_3N	Li_3N	1.4×10^{-10}	534	PFG NMR	94
$Li_{1.5}Al_{0.5}Ge_{1.5}(PO_4)_3$	NASICON	2.9×10^{-13}	311	PFG NMR	105
$Li_{1.2}Al_{0.2}Ti_{1.8}(PO_4)_3$	NASICON	1.5×10^{-12}	250	NMR relaxometry + PFG NMR	106
$Li_{6.6}La_3Zr_{1.6}Ta_{0.4}O_{12}$	Garnet	3.5×10^{-13}	353	PFG NMR	107
$Li_7La_3Zr_2O_{12}$	Garnet	1.8×10^{-18}	298	NMR relaxometry	108
$Li_7La_3Zr_2O_{12}$	Garnet	1.3×10^{-17}	325	NMR relaxometry	108
$Li_7La_3Zr_2O_{12}$	Garnet	3.3×10^{-14}	530	NMR relaxometry	108
Graphite (Stage I)	Graphite	$1–2\times10^{-15}$	298	NMR relaxometry	109
Li metal	bcc	5×10^{-15}	298	PFG NMR (extrapolated)	110
Li metal	bcc	1×10^{-12}	400	PFG NMR	110
$Li_4Ti_5O_{12}$	Spinel	3.2×10^{-15}	298	μ^+-SR	111,112
$Li_{5.7}Ti_5O_{12}$	Spinel	2.7×10^{-16}	298	NMR relaxometry	113
$LiTi_2O_4$	Spinel	3.6×10^{-15}	298	μ^+-SR	111
β-Li_2TiO_3	Li_2SnO_3	2×10^{-17}	433	NMR relaxometry	114
Li_xTiO_2 (x = 0.12) micro	Anatase component	4.7×10^{-16}	293	NMR relaxometry	115
Li_xTiO_2 (x = 0.12) micro	Li-titanate component	1.3×10^{-15}	293	NMR relaxometry	115
Li_xTiO_2 (x = 0.06) nano	Anatase component	1.9×10^{-16}	293	NMR relaxometry	116
Li_xTiO_2 (x = 0.06) nano	Li-titanate component	5.7×10^{-16}	293	NMR relaxometry	116
Li_xTiO_2 (x = 0.12) nano	Anatase component	1.1×10^{-16}	293	NMR relaxometry	116
Li_xTiO_2 (x = 0.12) nano	Li-titanate component	1.8×10^{-16}	293	NMR relaxometry	116
Li_xTiO_2 (x = 0.12) nano	Li-titanate component	4.9×10^{-16}	413	NMR relaxometry	116
$Li_{0.55}TiO_2$ nano	Li-titanate	1.7×10^{-16}	293	NMR relaxometry	116
$Li_{0.55}TiO_2$ nano	Li-titanate	4.6×10^{-16}	413	NMR relaxometry	116
$LiFePO_4$	Olivine	$1–6\times10^{-20}$	300	DFT/Tracer	49,50
$LiMn_2O_4$	Spinel	1×10^{-20}	350	NMR relaxometry	117
$LiMn_2O_4$	Spinel	1.8×10^{-16}	623	Tracer	118
$LiNbO_3$	$LiNbO_3$	7.5×10^{-15}	890	NMR relaxometry	119
Li_3NbO_4	Li_3NbO_4	4×10^{-21}	353	NMR relaxometry	120
Li_3NbO_4	Li_3NbO_4	1×10^{-16}	553	NMR relaxometry	120

Battery electrode materials are not typically amenable to the direct measurement of lithium diffusion via PFG NMR owing to rapid paramagnetic (T_1 and T_2) relaxation and slow diffusion. The weak paramagnetism (as indicated by the small NMR shift) and ultrafast diffusion of block-phase $Nb_{16}W_5O_{55}$ and bronze-like-phase $Nb_{18}W_{16}O_{93}$ enabled PFG NMR measurement, but only at slightly elevated temperature, where the T_2 relaxation time was increased. Room-temperature D_{Li} values were calculated by extrapolating from the elevated temperature data and assuming Arrhenius behaviour. Lithium diffusion measurements from NMR relaxometry are included here for several anode materials in which diffusion is too slow to be measured by PFG NMR[93]. cs, crystallographic shear; LISICON, lithium superionic conductor; NASICON, sodium superionic conductor; O1, layered structure with octahedral lithium and single-layer repeat stacking of transition-metal layers; PC, propylene carbonate; bcc, body-centred cubic; μ^+-SR, muon spin resonance; DFT, density functional theory.

© 2018 Springer Nature Limited. All rights reserved.

ARTICLE

https://doi.org/10.1038/s41586-018-0322-9

Evolution of a central neural circuit underlies *Drosophila* mate preferences

Laura F. Seeholzer[1], Max Seppo[1], David L. Stern[2] & Vanessa Ruta[1]*

Courtship rituals serve to reinforce reproductive barriers between closely related species. *Drosophila melanogaster* and *Drosophila simulans* exhibit reproductive isolation, owing in part to the fact that *D. melanogaster* females produce 7,11-heptacosadiene, a pheromone that promotes courtship in *D. melanogaster* males but suppresses courtship in *D. simulans* males. Here we compare pheromone-processing pathways in *D. melanogaster* and *D. simulans* males to define how these sister species endow 7,11-heptacosadiene with the opposite behavioural valence to underlie species discrimination. We show that males of both species detect 7,11-heptacosadiene using homologous peripheral sensory neurons, but this signal is differentially propagated to P1 neurons, which control courtship behaviour. A change in the balance of excitation and inhibition onto courtship-promoting neurons transforms an excitatory pheromonal cue in *D. melanogaster* into an inhibitory cue in *D. simulans*. Our results reveal how species-specific pheromone responses can emerge from conservation of peripheral detection mechanisms and diversification of central circuitry, and demonstrate how flexible nodes in neural circuits can contribute to behavioural evolution.

Animals display an extraordinary diversity of behaviour both within and between species. Although there is increasing insight into how learning and experience modify neural processing to produce variations in individual behaviour, far less is known about how evolution shapes neural circuitry to generate species-specific responses. Cross-species comparative studies have identified genetic loci that explain behavioural diversity[1–4], but only rarely have they examined the neural substrate upon which this genetic variation acts. Consequently, how the nervous system evolves to give rise to species-specific behaviours remains unclear.

As species diverge, their reproductive isolation is often reinforced by the development of behavioural differences that signify species identity and discourage interspecies courtship. The rapid evolution of courtship rituals therefore provides an entry point to examine the neural mechanisms that underlie behavioural divergence between closely related species. For example, *D. melanogaster* and *D. simulans* diverged 2–3 million years ago[5] (Fig. 1a) and, although these cosmopolitan species frequently encounter each other in the environment, they remain reproductively isolated and rarely mate. When copulation does occur, it results in inviable or sterile offspring. The avoidance of interspecies mating therefore benefits both species. Indeed, both *D. melanogaster* and *D. simulans* males are choosy in selecting mates and, when offered a choice, preferentially court conspecific females (Fig. 1b).

One mechanism for selective courtship is the use of sex- and species-specific pheromones that either promote the courtship of conspecific females or suppress the pursuit of inappropriate mates[6]. Pheromone production has rapidly diversified across drosophilids[7], such that *D. melanogaster* females predominantly produce 7,11-heptacosadiene (7,11-HD) on their cuticle whereas *D. simulans* females produce 7-tricosene (7-T), as do *D. simulans* and *D. melanogaster* males[6]. This pheromonal difference seems to be sufficient for species discrimination: perfuming *D. simulans* females with 7,11-HD renders them attractive to *D. melanogaster* males but unattractive to *D. simulans* males[8]. 7,11-HD is therefore detected by males of both species, but plays an opposing role in controlling their courtship decisions.

Species-specific pheromone responses could arise from the evolution of peripheral detection mechanisms or the central circuits that process pheromone signals to regulate courtship. To differentiate between these possibilities, we performed a direct comparison of the homologous pheromone pathways in *D. simulans* and *D. melanogaster* males. Here we demonstrate that species-specific responses to 7,11-HD emerge from the reweighting of excitatory and inhibitory inputs at a central node in the courtship circuit, highlighting how functional adaptations of central sensory processing pathways can lead to divergent behaviours.

Ppk23 mediates 7,11-HD detection

A critical step in *Drosophila* mate assessment occurs when a male taps the abdomen of another fly with his foreleg to taste their cuticular pheromones[9,10]. *D. melanogaster* and *D. simulans* males whose foreleg tarsi have been surgically removed still court vigorously (Extended Data Fig. 1a–c), but do so promiscuously (Fig. 1b). Therefore, although the detection of cuticular pheromones is not necessary for robust courtship[8], it is essential for species discrimination.

Several classes of gustatory sensory neurons on the *D. melanogaster* male foreleg detect pheromones to differentially regulate courtship. One heterogeneous sensory population expresses the Ppk23 DEG/ENaC channel: a subset of Ppk23$^+$ neurons detects female pheromones, including 7,11-HD, to promote courtship, and another subset detects male pheromones to inhibit courtship[11–13]. A smaller population of foreleg sensory neurons expresses the Gr32a receptor and detects 7-T to suppress inappropriate pursuit of *D. simulans* females and *D. melanogaster* males[10,14]. Although it is not known whether Gr32a and Ppk23 directly bind cuticular hydrocarbons, they serve as essential components of these pheromone transduction pathways[10–14]. We therefore investigated how Gr32a and Ppk23 shape mate preferences in *D. simulans* by using CRISPR–Cas9 genome editing to generate mutant alleles (Extended Data Fig. 1d). Both *Gr32a* and *ppk23* mutant males pursued conspecific females with the same vigour as wild-type *D. simulans* males (Fig. 1c), suggesting that either these receptors do not contribute

[1]Laboratory of Neurophysiology and Behavior, The Rockefeller University, New York, NY, USA. [2]Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, VA, USA.
*e-mail: ruta@rockefeller.edu

© 2018 Springer Nature Limited. All rights reserved.

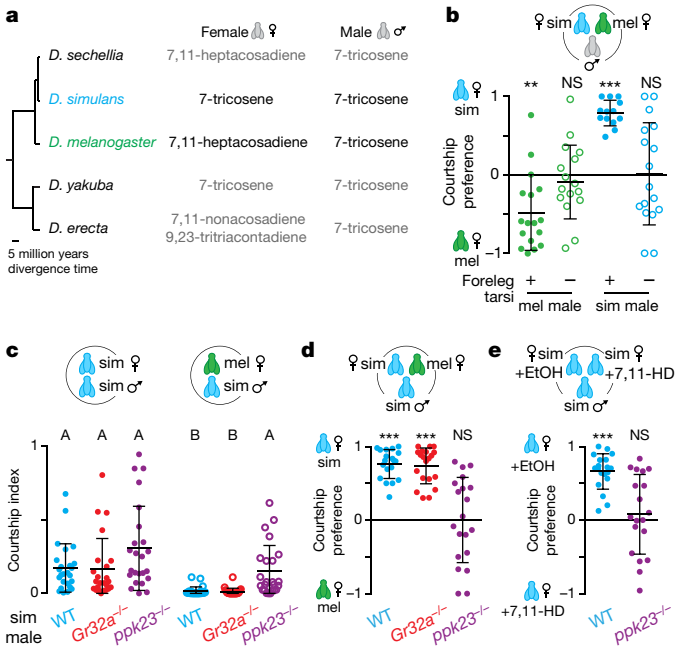


Fig. 1 | Pheromone regulation of *D. simulans* courtship. a, Predominant cuticular hydrocarbons of related species. **b**, Courtship preferences of *D. melanogaster* (mel) and *D. simulans* (sim) males with foreleg tarsi intact (+) or removed (−). **c**, **d**, Courtship indices (**c**) and preference indices (**d**) of wild-type (WT), *Gr32a*$^{-/-}$ and *ppk23*$^{-/-}$ *D. simulans* males offered *D. simulans* and/or *D. melanogaster* females. **e**, Preference indices of wild-type or *ppk23*$^{-/-}$ *D. simulans* males offered *D. simulans* females perfumed with ethanol (EtOH) or 7,11-HD. Statistical tests used were a one-sample *t*-test (**b**, **d**, **e**) and a Kruskal–Wallis test (**c**). $^{**}P < 0.01$, $^{***}P < 0.001$, NS, not significant (**b**, **d**, **e**); different letters mark significant differences in courtship by Dunn's multiple comparison test (**c**). Data are mean and s.d., with individual data points shown. See Supplementary Table 1 for details of statistical analyses.

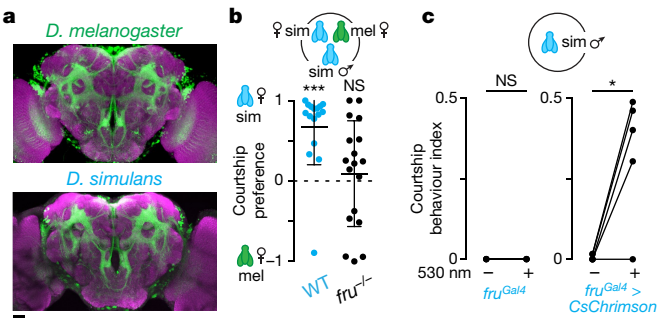


Fig. 2 | Fruitless has a conserved role in male courtship. a, Fru^{+} neurons (green) with neuropil counterstain (magenta) in brains of *D. melanogaster* and *D. simulans* males. Scale bar, 30 μm. **b**, Courtship preference indices of wild-type (WT) and *fru*$^{-/-}$ *D. simulans* males. **c**, Fraction of time that solitary males displayed courtship behaviours with (right) or without (left) optogenetic stimulation of Fru^{+} neurons. Statistical tests used were a one-sample *t*-test (**b**) and a Wilcoxon matched-pairs signed-rank test (**c**). $^{*}P < 0.05$, $^{***}P < 0.001$. Data are mean and s.d., with individual data points shown. See Supplementary Table 1 for details of statistical analyses.

to the detection of 7-T in *D. simulans*, or that 7-T does not play a role in promoting male courtship[15], despite being the predominant cuticular pheromone on *D. simulans* females[6]. *D. simulans Gr32a* mutant males also did not court *D. melanogaster* females or females of more distant species (Fig. 1c, d, Extended Data Fig. 1e, f). Therefore, contrary to its role in mediating courtship suppression in *D. melanogaster*[10,14], Gr32a does not seem to influence mate choices in *D. simulans*.

By contrast, *ppk23* mutants pursued *D. melanogaster* females and other drosophilids carrying diene pheromones with the same intensity as they courted *D. simulans* females (Fig. 1c, Extended Data Fig. 1e), and were unable to differentiate between *D. melanogaster* and *D. simulans* females in preference assays (Fig. 1d, Extended Data Fig. 1f). Moreover, *ppk23* mutants were not deterred from courting *D. simulans* females that were perfumed with 7,11-HD (Fig. 1e, Extended Data Fig. 1g), suggesting that their promiscuous courtship reflects an inability to detect the *D. melanogaster* pheromone. As such, males of both species rely on *ppk23* to detect 7,11-HD, but detection of this pheromone initiates opposing behaviours in the two species—promoting courtship in *D. melanogaster* while suppressing courtship in *D. simulans*. We therefore developed genetic tools in *D. simulans* to examine the sensory neurons in which Ppk23 is expressed and the downstream circuits that they activate to identify the neural adaptations that contribute to species-specific pheromone responses.

fruitless specifies courtship circuitry

In *D. melanogaster*, the male-specific isoform of the Fruitless transcription factor (Fru^{M})[16,17] marks the majority of Ppk23^{+} sensory neurons in the male foreleg[11–13] along with the neural circuitry mediating most of the components of male courtship, from sensory detection to motor implementation. To gain genetic access to the repertoire of Fru^{+} neurons in *D. simulans*, including those that detect and process 7,11-HD, we integrated either the GFP or Gal4 coding sequence into the first intron of the *fru* locus (Extended Data Fig. 2a). We observed that, in both species, *fru* marks a similar ensemble of neurons distributed throughout the male nervous system, with comparable innervation patterns evident in most brain neuropils (Fig. 2a, Extended Data Fig. 2b–g).

Given the anatomic similarity of Fru^{+} neurons, we wanted to confirm the evolutionary conservation of *fru* as a master regulator of male courtship behaviours[16–20]. We found that, as in *D. melanogaster*, mutation of *fru*^{M} resulted in *D. simulans* males that exhibited aberrant mate preferences, whereas the optogenetic activation of Fru^{+} neurons triggered multiple components of the courtship ritual in an isolated male (Fig. 2b, c, Extended Data Fig. 2h, i). Therefore, in both species, *fru* marks circuits that specify male courtship towards appropriate sexual partners, providing an inroad to trace and compare the neural pathways that underlie mate discrimination, from the sensory periphery to higher brain centres.

Conserved peripheral sensory responses

To investigate the pheromone tuning of the Ppk23^{+} sensory population in *D. melanogaster* and *D. simulans* males, we expressed the Ca^{2+} indicator GCaMP6s in Fru^{+} neurons and monitored the aggregate responses of foreleg sensory afferents in the ventral nerve cord of a male as his tarsus contacted the abdomen of a female (Fig. 3a). We found that the Fru^{+} sensory neurons of both *D. melanogaster* and *D. simulans* males exhibited comparable pheromone tuning, responding robustly to the taste of a *D. melanogaster* female but weakly to that of a *D. simulans* female (Fig. 3b, c). In males of both species, sensory neuron responses were strongly attenuated in *ppk23* mutants (Fig. 3b, c, Extended Data Fig. 3e), verifying that Ppk23 plays a conserved and essential role in pheromone detection.

To compare the distribution of 7,11-HD-responsive sensory neurons across species directly, we generated a *ppk23-Gal4* construct in *D. simulans*, taking advantage of the fact that, in *D. melanogaster*, the *ppk23* promoter faithfully reproduces endogenous channel expression[12]. We found that the *ppk23* promoter from either species drove expression in a comparable number of sensory neurons in the male foreleg, whose axons exhibited a similar sexually dimorphic projection pattern within the ventral nerve cord (Extended Data Fig. 3a–d). Imaging the aggregate activity of these Ppk23^{+} sensory afferents revealed equivalent pheromone tuning across males of both species, with significantly stronger responses to the taste of a *D. melanogaster* female than to that of a *D. simulans* female (Fig. 3d, Extended Data Fig. 3e). Moreover, we observed no difference in the distribution or magnitude of functional responses in Ppk23^{+} soma when we stimulated individual

© 2018 Springer Nature Limited. All rights reserved.

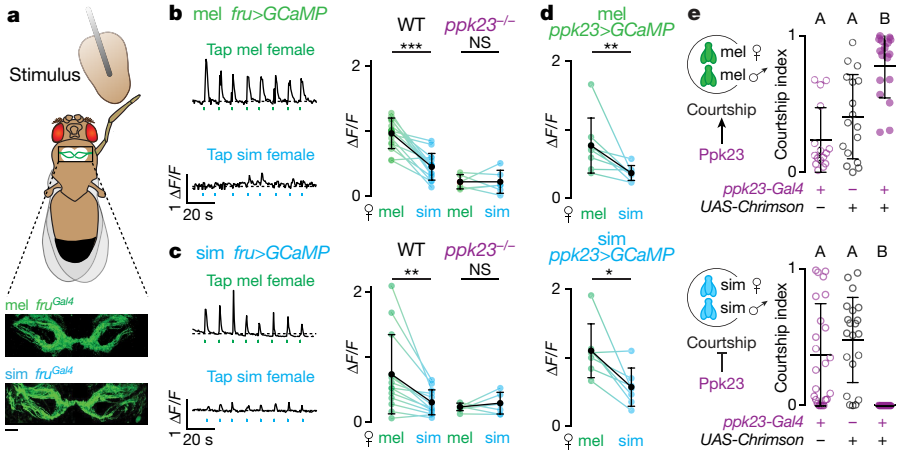


Fig. 3 | Conserved pheromonal tuning of Ppk23$^+$ and Fru$^+$ foreleg sensory neurons. a, Schematic of ventral nerve cord preparation used for imaging (top) and Fru$^+$ foreleg sensory neurons expressing GCaMP (bottom). Scale bar, 10 μm. **b**, **c**, Functional responses of Fru$^+$ foreleg afferents in *D. melanogaster* (**b**) or *D. simulans* (**c**) males evoked by the taste of a *D. melanogaster* or *D. simulans* female. Representative activity traces with time of taps indicated by coloured tick marks (left) and average normalized fluorescence ($\Delta F/F$) in wild-type (WT) and *ppk23*$^{-/-}$ males (right). **d**, Functional responses (average $\Delta F/F$) of Ppk23$^+$ foreleg afferents evoked by the taste of a *D. melanogaster* or *D. simulans* female. **e**, Courtship indices towards conspecific females during optogenetic stimulation of Ppk23$^+$ neurons in *D. melanogaster* (top) and *D. simulans* males (bottom). Statistical tests used were a Wilcoxon matched-pairs signed-rank test (**b**–**d**) and a Kruskal–Wallis test (**e**). $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ (**b**–**d**); different letters mark significant differences in courtship indices by Dunn's multiple comparison test (**e**). Data are mean and s.d., with individual data points shown. See Supplementary Table 1 for details of statistical analyses.

sensory bristles on the foreleg with synthetic 7,11-HD (Extended Data Fig. 3f). As in *D. melanogaster*[12,21], Ppk23$^+$ soma were paired beneath a sensory bristle, and only one soma of each pair responded to 7,11-HD. Somatic responses to synthetic 7-T were negligible in both species (data not shown), mirroring the weak responses evoked at the population level by the taste of a *D. simulans* female (Fig. 3d).

Together, these experiments demonstrate that a quantitatively and qualitatively similar population of Ppk23$^+$ sensory neurons is tuned to 7,11-HD in both *D. melanogaster* and *D. simulans* males. To assess whether the activation of Ppk23$^+$ sensory neurons could replicate the opposing courtship behaviours elicited by 7,11-HD, we expressed CsChrimson in this sensory population in males of both species and examined how optogenetic activation influenced courtship of a conspecific female. Whereas optogenetic stimulation of Ppk23$^+$ sensory neurons in *D. melanogaster* males drove increased courtship[12,13], it inhibited courtship towards an otherwise attractive conspecific female in *D. simulans* males (Fig. 3e), replicating the courtship suppression that results from perfuming a *D. simulans* female with 7,11-HD. Therefore, activation of the homologous Ppk23$^+$ sensory neuron population is sufficient to drive opposing behavioural responses in *D. melanogaster* and *D. simulans* males, suggesting that differences must exist in the downstream circuits that link pheromone detection to courtship decisions.

Divergent responses in central circuits

In *D. melanogaster*, male-specific P1 neurons form a central node in the Fru$^+$ circuitry that integrates input from multisensory pathways to represent the suitability of a potential mate and triggers the initiation of courtship[21–26]. Anatomic labelling revealed that, in both species, P1 neurons exhibit rich projections in the lateral protocerebral complex (LPC, Fig. 4a), a sexually dimorphic Fru$^+$ neuropil that receives input from multiple sensory processing pathways and outputs to descending neurons that drive the component behaviours of male courtship[21,25,27].

To test whether P1 neurons play a conserved role in regulating courtship across species, we introduced a transcriptional enhancer (*R71G01-Gal4*) that labels this neural population in *D. melanogaster*[23] into the *D. simulans* genome and used it to drive expression of CsChrimson (Extended Data Fig. 4a). Optogenetic activation of P1 neurons in *D. simulans* males strongly enhanced courtship of conspecific females and drove almost incessant courtship of inappropriate targets, including *D. melanogaster* females and a small rotating magnet (Fig. 4b, Supplementary Video 1). Pursuit of the magnet was significantly reduced when it was stationary or moving slowly, highlighting the importance of motion for vigorous courtship[23,26] (Extended Data Fig. 4c). Courtship towards all targets remained increased after stimulation, indicating that transient activation of P1 neurons is sufficient to trigger an enduring state of sexual arousal across species[24–26,28] (Fig. 4b, Extended Data Fig. 4b–f).

Notably, titration of the stimulating light revealed that evoked courtship in *D. simulans* males was significantly weaker towards *D. melanogaster* females than *D. simulans* females (Fig. 4c), raising the possibility that 7,11-HD may suppress courtship by countering the excitation of P1 neurons. To compare how pheromone signals are propagated from the periphery to P1 neurons and other central Fru$^+$ populations, we monitored responses either in the LPC or in the P1 neurons in a tethered male as he tapped the abdomen of a target fly with his foreleg (Fig. 4d, Extended Data Fig. 4g). In *D. melanogaster* males, robust responses were evoked in the LPC by the taste of a *D. melanogaster* female, but not a *D. simulans* female, reflecting strong excitation of P1 neurons by the pheromones of an appropriate mate[25,29] (Fig. 4e, g, Extended Data Fig. 4h, Supplementary Videos 2, 3). By contrast, in *D. simulans* males, neither the P1 neurons nor any other Fru$^+$ neural populations in the LPC were activated in response to the taste of a *D. simulans* female (Fig. 4f, Supplementary Video 4), which is consistent with behavioural evidence that contact pheromones are not necessary to promote *D. simulans* courtship[8,15]. Neurons in the LPC of *D. simulans* males were, however, weakly activated by the taste of a *D. melanogaster* female, but these signals failed to propagate to the P1 neurons (Fig. 4f, h, Supplementary Video 5). Responses were not apparent in *ppk23* mutants, verifying that pheromone signalling to the LPC relies on Ppk23$^+$ sensory pathways (Extended Data Fig. 4i, j). Opposing behavioural responses to 7,11-HD in the two species therefore appear to be mirrored by divergent P1 neuron excitation.

Species-specific central circuit changes

In *D. melanogaster* and *D. simulans* males, equivalent pheromone responses at the sensory periphery are translated into differential excitation of P1 neurons, suggesting that the ascending pathways that convey pheromone signals from the forelegs to the LPC have diverged. In *D. melanogaster*, 7,11-HD signals are transmitted from Ppk23$^+$ sensory neurons to vAB3 neurons, whose dendrites reside in the ventral nerve cord and whose axons project to the LPC, providing a direct excitatory route to the P1 neurons[25] (Fig. 5a). vAB3 neurons

© 2018 Springer Nature Limited. All rights reserved.

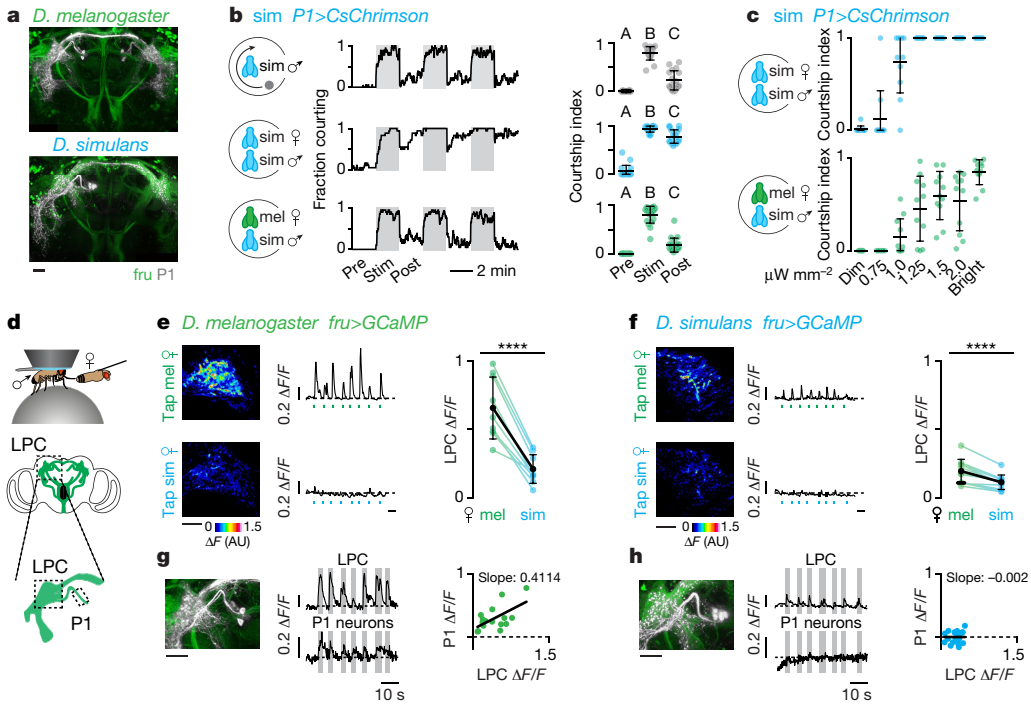


Fig. 4 | Divergent pheromone responses in courtship-promoting P1 neurons. a, P1 neurons (grey) innervate the Fru$^+$ (green) LPC. **b**, **c**, Optogenetic stimulation of P1 neurons in *D. simulans* males. **b**, Courtship towards a rotating magnet (top), *D. simulans* female (middle), or *D. melanogaster* female (bottom). Fraction of flies courting (left; grey boxes indicate illumination with bright light) and courtship indices pre-, during, and post-stimulation (right). **c**, Courtship indices of *D. simulans* males towards *D. simulans* (top) and *D. melanogaster* (bottom) females in dim or bright white light or the indicated illumination intensity at 627 nm. **d**, Experimental set-up for in vivo functional GCaMP imaging (top) and schematic of Fru$^+$ neurons (bottom; the regions imaged are outlined). **e**, **f**, Functional responses of Fru$^+$ neurons in the LPC of *D. melanogaster* (**e**) and *D. simulans* (**f**) males evoked by tapping *D. melanogaster* and *D. simulans* females. Representative GCaMP fluorescence increase upon stimulation (left). Representative activity traces with time of stimulating taps indicated (middle). Average $\Delta F/F$ for individual males (right). AU, arbitrary units. **g**, **h**, Simultaneously recorded activity of LPC and P1 neurons in *D. melanogaster* (**g**) and *D. simulans* (**h**) males evoked by the taste of a *D. melanogaster* female. Magnified P1 neuron anatomy (left; same images as in **a**). Representative activity traces (middle; grey bars indicate taps) and relationship between responses in LPC and P1 neurons evoked by individual taps (right). Statistical tests used were a Kruskal–Wallis test (**b**), a paired *t*-test (**e**, **f**) and linear regression (**g**, **h**). Different letters mark significant differences in courtship indices by Dunn's multiple comparison test (**b**); $^{****}P < 0.0001$ (**e**, **f**); $R^2 = 0.3575$ (**g**) and $R^2 = 0.0275$ (**h**). Data are mean and s.d., with individual data points shown. Scale bars for anatomic images, 10 μm. See Supplementary Table 1 for details of statistical analyses.

also extend collaterals into the suboesophageal zone (SEZ), where they synapse onto GABAergic (γ-aminobutyric acid-releasing) mAL interneurons[25]. mAL axons arborize extensively in the LPC and provide inhibitory input onto P1 neurons, forming a feedforward inhibitory circuit motif that tempers P1 neuron excitation and regulates the gain of pheromone responses[21,25]. In *D. melanogaster*, P1 neurons thus receive excitatory and inhibitory input even in response to the taste of a conspecific female, with 7,11-HD evoking net excitation to trigger courtship. Anatomic labelling of *D. simulans* vAB3 and mAL neurons revealed broadly conserved projection patterns within the LPC (Fig. 5b, Extended Data Fig. 5), indicating that they remain anatomically poised to synapse onto P1 neurons.

Given the structural conservation of these ascending pathways, we considered whether there might be functional differences in how pheromone signals are transmitted through this circuit to generate divergent P1 neuron responses. Functional imaging revealed that vAB3 neurons were similarly tuned to pheromones across species, with robust responses elicited only by the taste of a *D. melanogaster* female, and not a *D. simulans* female (Fig. 5c, Extended Data Fig. 6a–d). Moreover, in both species, vAB3 pheromone responses were lost in *ppk23* mutants (Extended Data Fig. 6b–d). Therefore, although Ppk23$^+$ sensory neurons drive opposing courtship behaviours in *D. melanogaster* and *D. simulans* males, they nevertheless elicit comparable pheromone responses in the homologous ascending pathways in the brain.

To compare the role of mAL neurons across species directly, we introduced a genetic driver (*R25E04-Gal4*) that labels mAL neurons[25] into *D. simulans* (Extended Data Fig. 6e). Optogenetic activation of mAL neurons in *D. simulans* males strongly attenuated courtship (Extended Data Fig. 6f), replicating the robust courtship suppression mediated by mAL neurons in *D. melanogaster*[21] and confirming that this population inhibits neurons in the LPC in both species. Functional imaging revealed that mAL neurons were equivalently excited by the taste of a *D. melanogaster* female in both *D. melanogaster* and *D. simulans* males (Fig. 5d, Extended Data Fig. 6g). Therefore, in both species vAB3 and mAL pathways are similarly activated by *D. melanogaster* female pheromones, suggesting that alterations in the strength of their signalling to P1 neurons might underlie the emergence of species-specific mate preferences, such that 7,11-HD can evoke excitation of P1 neurons only in *D. melanogaster* males to initiate courtship of a *D. melanogaster* female.

To examine this possibility, we directly stimulated vAB3 neurons through iontophoresis of acetylcholine onto their dendrites within the ventral nerve cord and performed multi-plane functional imaging to visualize activated Fru$^+$ neurons in the brain (Fig. 5e, Extended Data Fig. 7a, b). Although this stimulation elicited equivalent responses in the vAB3 and mAL neurons of both species, P1 neurons and other neurons of the LPC were excited only in *D. melanogaster* and not in *D. simulans* males (Fig. 5e, f, Extended Data Fig. 7d–g), mirroring the differential propagation of pheromone signals through this pathway in vivo. In both species, the responses of all Fru$^+$ neurons were lost after we severed the vAB3 axons with a two-photon laser, verifying that vAB3 neurons mediate these divergent activity patterns (Extended Data Fig. 7c).

To assess whether mAL-mediated inhibition could counter vAB3 excitation to suppress P1 neurons in *D. simulans*, we used a two-photon laser to sever the mAL axonal tract and prevent the transmission

© 2018 Springer Nature Limited. All rights reserved.

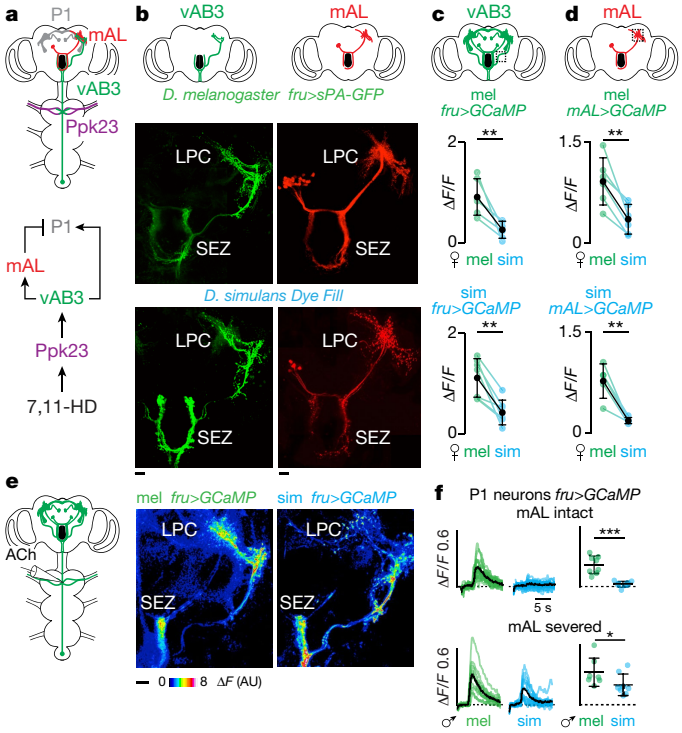


Fig. 5 | Differential propagation of ascending pheromone signals to P1 neurons. a, Schematic of the Fru$^+$ circuit that processes 7,11-HD. **b**, Anatomy of vAB3 and mAL neurons in *D. melanogaster* and *D. simulans* with the LPC and SEZ marked. Scale bars, 10 μm. **c**, **d**, Average $\Delta F/F$ for individual males evoked by *D. melanogaster* and *D. simulans* females in vAB3 (**c**) and mAL (**d**) neurons. **e**, Representative GCaMP fluorescence increase in Fru$^+$ neurons evoked by direct vAB3 stimulation with acetylcholine (ACh). **f**, P1 neuron responses to vAB3 stimulation before (top) and after (bottom) mAL severing in *D. melanogaster* and *D. simulans* males. Left, coloured lines depict individual stimulations and black lines show the average; right, the graph plots the peak $\Delta F/F$ per fly. Scale bars are the same for P1 activity traces and graphs. Statistical tests used are a paired *t*-test (**c**, **d**) and an unpaired *t*-test with Welch correction (**f**). $^*P < 0.05$; $^{**}P < 0.01$, $^{***}P < 0.001$. Data are mean and s.d., with individual data points shown. See Supplementary Table 1 for details of statistical analyses.

of their GABAergic signal to the LPC. After the severing of mAL, we found that P1 neurons could be excited by vAB3 stimulation in *D. simulans* males, although to a lower level than in *D. melanogaster* males (Fig. 5f, Extended Data Fig. 7g). Pharmacological weakening of inhibition by injection of the GABA-receptor antagonist, picrotoxin, into the LPC similarly enhanced pheromone responses in vivo, unmasking excitation specifically in response to the taste of a *D. melanogaster* female (Extended Data Fig. 6h). These results suggest that mAL-mediated inhibition antagonizes vAB3 excitation to fully suppress P1 neuron responses in *D. simulans* males, but not in *D. melanogaster* males, revealing how alterations in excitatory and inhibitory input to this population may generate divergent responses to the same pheromone cue.

Discussion

The sensory periphery has been proposed to be the most evolutionarily labile element of the nervous system[30–32], as changes in the expression or tuning of sensory receptors can allow for the emergence of species-specific behaviours without necessitating potentially more complex developmental rewiring of central pathways in the brain. By contrast, our results suggest that species-specific behavioural responses to 7,11-HD arise through functional alterations in how pheromone signals are propagated through a structurally conserved central circuit that consists of parallel excitatory and feedforward inhibitory branches. By reweighting the balance of excitatory vAB3 and inhibitory mAL signalling to P1 neurons, 7,11-HD is transformed from an excitatory signal that promotes courtship in *D. melanogaster* males into an inhibitory signal that suppresses courtship in *D. simulans* males. Although our analysis highlights the importance of this pheromone pathway in shaping species-specific mate preferences, we cannot exclude the possibility that additional inputs to P1 neurons or other targets of mAL and vAB3 neurons also contribute to divergent courtship decisions. Nevertheless, our data suggest that the branched architecture of pheromone-processing pathways serves as a substrate for the evolution of mate preferences, pointing to the potential existence of favourable sites within neural circuits to instantiate adaptive behavioural changes, analogous to how specific nodes within developmental regulatory networks contribute to morphological diversity[33].

The conserved tuning of Ppk23$^+$ neurons suggests that *D. simulans* males dedicate this sensory pathway to the detection of *D. melanogaster* female pheromones, rather than sensing the chemical cues carried by conspecific females. Consistent with this idea, although P1 neurons are sufficient to elicit courtship in *D. simulans* males, they are not excited by the taste of a *D. simulans* female. Given that *D. simulans* cuticular pheromones are sexually monomorphic and offer ambiguous signals for mate recognition, males probably rely on additional sensory inputs for their arousal. Indeed, the fervent courtship exhibited by *D. simulans ppk23* mutants towards females of different species demonstrates that *D. simulans* males can be aroused in the absence of any species-specific excitatory cue. Dienes, such as 7,11-HD, actively suppress this arousal, presumably through recruitment of strong mAL-mediated inhibition via Ppk23$^+$ pathways. Together, these observations reinforce the notion that pheromone communication in *Drosophila* serves to focus a male's desire, such that flies lacking cuticular pheromones can be inherently attractive and appropriate mate choices are honed by specific inhibitory chemical cues[8,10,14].

Peripheral adaptations are likely to play an important role in the evolution of novel chemical sensitivities[2,31,32]. In this case, however, preserving the sensory periphery while varying central circuits provides a mechanism to alter the behavioural valence of a single chemical cue. As *D. melanogaster* and *D. simulans* diverged, their reproductive isolation was probably strengthened by the ability of both species to detect the same pheromone but assign it a different meaning through these central circuit modifications. Notably, in *D. melanogaster*, P1 neuron excitability is regulated by the social history of a male[24,34]. This suggests that both experience-dependent and evolutionary adaptations may act on the same neural substrate to modify sensory integration and mate choices, similar to how phenotypic plasticity may facilitate morphological evolution[35]. Therefore, functional reweighting of sensory inputs at flexible nodes in the nervous system, shaped by evolutionary selection or individual experience, may allow for alternative behavioural responses to the same sensory signal.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0322-9.

Received: 27 March 2017; Accepted: 5 June 2018;
Published online 11 July 2018.

1. Bendesky, A. et al. The genetic basis of parental care evolution in monogamous mice. *Nature* **544**, 434–439 (2017).
2. McGrath, P. T. et al. Parallel evolution of domesticated *Caenorhabditis* species targets pheromone receptor genes. *Nature* **477**, 321–325 (2011).
3. Weber, J. N., Peterson, B. K. & Hoekstra, H. E. Discrete genetic modules are responsible for complex burrow evolution in *Peromyscus* mice. *Nature* **493**, 402–405 (2013).
4. Ding, Y., Berrocal, A., Morita, T., Longden, K. D. & Stern, D. L. Natural courtship song variation caused by an intronic retroelement in an ion channel gene. *Nature* **536**, 329–332 (2016).
5. Hey, J. & Kliman, R. M. Population genetics and phylogenetics of DNA sequence variation at multiple loci within the *Drosophila melanogaster* species complex. *Mol. Biol. Evol.* **10**, 804–822 (1993).

© 2018 Springer Nature Limited. All rights reserved.

6. Jallon, J.-M. & David, J. R. Variation in cuticular hydrocarbons among the eight species of the *Drosophila melanogaster* subgroup. *Evolution* **41**, 294–302 (1987).
7. Shirangi, T. R., Dufour, H. D., Williams, T. M. & Carroll, S. B. Rapid evolution of sex pheromone-producing enzyme expression in *Drosophila*. *PLoS Biol*. **7**, e1000168 (2009).
8. Billeter, J.-C., Atallah, J., Krupp, J. J., Millar, J. G. & Levine, J. D. Specialized cells tag sexual and species identity in *Drosophila melanogaster*. *Nature* **461**, 987–991 (2009).
9. Bastock, M. & Manning, A. The courtship of *Drosophila melanogaster*. *Behaviour* **8**, 85–110 (1955).
10. Fan, P. et al. Genetic and neural mechanisms that inhibit *Drosophila* from mating with other species. *Cell* **154**, 89–102 (2013).
11. Lu, B., LaMora, A., Sun, Y., Welsh, M. J. & Ben-Shahar, Y. *ppk23*-dependent chemosensory functions contribute to courtship behavior in *Drosophila melanogaster*. *PLoS Genet.* **8**, e1002587 (2012).
12. Thistle, R., Cameron, P., Ghorayshi, A., Dennison, L. & Scott, K. Contact chemoreceptors mediate male–male repulsion and male–female attraction during *Drosophila* courtship. *Cell* **149**, 1140–1151 (2012).
13. Toda, H., Zhao, X. & Dickson, B. J. The *Drosophila* female aphrodisiac pheromone activates *ppk23*$^{+}$ sensory neurons to elicit male courtship behavior. *Cell Rep.* **1**, 599–607 (2012).
14. Miyamoto, T. & Amrein, H. Suppression of male courtship by a *Drosophila* pheromone receptor. *Nat. Neurosci.* **11**, 874–876 (2008).
15. Coyne, J. A., Crittenden, A. P., Mah, K. & Maht, K. Genetics of a pheromonal difference contributing to reproductive isolation in *Drosophila*. *Science* **265**, 1461–1464 (1994).
16. Stockinger, P., Kvitsiani, D., Rotkopf, S., Tirián, L. & Dickson, B. J. Neural circuitry that governs *Drosophila* male courtship behavior. *Cell* **121**, 795–807 (2005).
17. Manoli, D. S. et al. Male-specific *fruitless* specifies the neural substrates of *Drosophila* courtship behaviour. *Nature* **436**, 395–400 (2005).
18. Demir, E. & Dickson, B. J. *fruitless* splicing specifies male courtship behavior in *Drosophila*. *Cell* **121**, 785–794 (2005).
19. Cande, J., Stern, D. L., Morita, T., Prud'homme, B. & Gompel, N. Looking under the lamp post: neither *fruitless* nor *doublesex* has evolved to generate divergent male courtship in *Drosophila*. *Cell Rep.* **8**, 363–370 (2014).
20. Tanaka, R., Higuchi, T., Kohatsu, S., Sato, K. & Yamamoto, D. Optogenetic activation of the *fruitless*-labeled circuitry in *Drosophila subobscura* males induces mating motor acts. *J. Neurosci.* **37**, 11662–11674 (2017).
21. Kallman, B. R., Kim, H. & Scott, K. Excitation and inhibition onto central courtship neurons biases *Drosophila* mate choice. *eLife* **4**, e11188 (2015).
22. von Philipsborn, A. C. et al. Neuronal control of *Drosophila* courtship song. *Neuron* **69**, 509–522 (2011).
23. Pan, Y., Meissner, G. W. & Baker, B. S. Joint control of *Drosophila* male courtship behavior by motion cues and activation of male-specific P1 neurons. *Proc. Natl Acad. Sci. USA* **109**, 10065–10070 (2012).
24. Inagaki, H. K. et al. Optogenetic control of *Drosophila* using a red-shifted channelrhodopsin reveals experience-dependent influences on courtship. *Nat. Methods* **11**, 325–332 (2014).
25. Clowney, E. J., Iguchi, S., Bussell, J. J., Scheer, E. & Ruta, V. Multimodal chemosensory circuits controlling male courtship in *Drosophila*. *Neuron* **87**, 1036–1049 (2015).
26. Kohatsu, S. & Yamamoto, D. Visually induced initiation of *Drosophila* innate courtship-like following pursuit is mediated by central excitatory state. *Nat. Commun.* **6**, 6457 (2015).
27. Yu, J. Y., Kanai, M. I., Demir, E., Jefferis, G. S. X. E. & Dickson, B. J. Cellular organization of the neural circuit that drives *Drosophila* courtship behavior. *Curr. Biol.* **20**, 1602–1614 (2010).
28. Ding, Y. et al. Neural changes underlying rapid fly song evolution. Preprint at https://doi.org/10.1101/238147 (2017).
29. Kohatsu, S., Koganezawa, M. & Yamamoto, D. Female contact activates male-specific interneurons that trigger stereotypic courtship behavior in *Drosophila*. *Neuron* **69**, 498–508 (2011).
30. Tierney, A. J. Evolutionary implications of neural circuit structure and function. *Behav. Processes* **35**, 173–182 (1995).
31. Cande, J., Prud'homme, B. & Gompel, N. Smells like evolution: the role of chemoreceptor evolution in behavioral change. *Curr. Opin. Neurobiol.* **23**, 152–158 (2013).
32. Bendesky, A. & Bargmann, C. I. Genetic contributions to behavioural diversity at the gene–environment interface. *Nat. Rev. Genet.* **12**, 809–820 (2011).
33. Stern, D. L. The genetic causes of convergent evolution. *Nat. Rev. Genet.* **14**, 751–764 (2013).
34. Zhang, S. X., Rogulja, D. & Crickmore, M. A. Dopaminergic circuitry underlying mating drive. *Neuron* **91**, 168–181 (2016).
35. West-Eberhard, M. J. Developmental plasticity and the origin of species differences. *Proc. Natl Acad. Sci. USA* **102**, (Suppl 1), 6543–6549 (2005).

Acknowledgements We thank Y. Ding for *D. simulans* CsChrimson flies; P. Pires-Mussells and J. Marquina-Solis for assistance on initial experiments; E. Clowney, B. Matthews, K. Kistler, J. Petrillo and G. Maimon for technical advice; and E. Clowney, S. Datta, B. Noro, L. Vosshall, S. Shaham, C. McBride and members of the Ruta laboratory for discussion and comments on the manuscript. This work was supported by the New York Stem Cell Foundation, the Pew Foundation, the McKnight Foundation, the Irma T. Hirschl Foundation, the Alfred P. Sloan Foundation and the National Institutes of Health (DP2 NS0879422013) to V.R. and a National Science Foundation and Kavli Fellowship to L.F.S.

Reviewer information *Nature* thanks B. Prud'homme and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions L.F.S. and V.R. conceived and designed the study. L.F.S. and M.S. performed behavioural experiments and generated receptor mutants and *fru* alleles in *D. simulans*. D.L.S. generated *R71G01, R25E04* and GCaMP strains in *D. simulans*. M.S. performed immunohistochemistry experiments. L.F.S. performed functional and anatomic tracing experiments. L.F.S. and V.R. analysed data and wrote the manuscript with input from all the authors.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0322-9.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0322-9.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to V.R.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Data reporting. Preliminary experiments were used to assess variance and determine adequate sample sizes in advance of conducting the experiment. We used random.org/sequences to randomize the order of all behaviour experiments. To control for potential variations in experimental conditions across days, we collected a similar sample size for each variable, every day the experiment was conducted. For functional experiments, we interleaved genotypes and female stimuli when applicable. All behaviour experiments were conducted with the experimenter blinded to the genotype or condition of any male or female fly that was a variable in a given experiment. The experimenter was unblinded only after analysis of the assay. The experimenter was not blinded to the genotype of males used in functional imaging assays.

Fly stocks and husbandry. Flies were housed under standard conditions at 25 °C under a 12 h light:12 h dark cycle. Fly stocks and sources were as follows: *Drosophila melanogaster* Canton-S, *20xUAS-IVS-GCaMP6s* (Bloomington; 42746, 42749), *UAS-mCD8::GFP* (5130, 5137), *LexAop-GCaMP6s* (53747), *10xUAS-IVS-myr::tdTomato* (32222), *R71G01-Gal4* (39599), *AbdB-Gal4* (55848) *R25E04-Gal4* (49125), and *20xUAS-IVS-CsChrimson.mVenus* (55134) were obtained from the Bloomington Stock Center. *D. sechellia* (14021-0248.25) and *D. erecta* (14021-0224.01) were obtained from the UCSD Stock Center. The following were obtained as indicated: *D. ananassae* and *D. simulans* (Ruta laboratory), *D. simulans attP2039*[36] (Y. Ding and D. Stern, Janelia Research Campus); *SplitP1-Gal4*[24] (D. Anderson, Caltech); fru^{LexA}[37] and fru^{Gal416} (B. Dickson, HHMI/Janelia Farm Research Campus); *D. melanogaster ppk23-Gal4*[11] (K. Scott, UC Berkeley); *LexAop-myr::tdTomato* (C. Mendes, Columbia University); *UAS-SPA-t2a-SPA-GFP*[25] (Ruta laboratory). Supplementary Table 1 provides detailed descriptions of genotypes used in each experiment.

Courtship behaviour assays and analysis. To standardize fly size and life history across trials, all flies used for behavioural assays were reared in food vials at a low density (3 females and 3 males seeded into vial). All experimental flies were male and all stimulus flies were virgin females unless otherwise noted. Males for all assays were collected as virgins, placed in individual food vials ($d = 3$ cm, $h = 9.5$ cm) and housed in isolation for 3–6 days. Males were added to behavioural assays by direct aspiration from the food vial without ice or CO_2 anaesthetization, except for the tarsi ablation experiments in which males were ice-anaesthetized. Virgin females were group-housed in food vials and aged 3–6 days. All behavioural assays were conducted at Zeitgeber 0 to 3 h except for assays using flies reared in the dark. All behavioural assays were conducted in a heated, humidified room (25 °C, 46% relative humidity) on a back-lit surface (Logan Electric Slim Edge-Light Pad A-5A, 5400K, 6 klx). For all preference assays, only males who spent more than 5% of the time courting (>30 s of total courtship) were included in the analysis. Courtship behaviours included in the analysis were singing, tapping, licking, orienting, abdomen bending and chasing. QuickTime7 software was used when analysing courtship videos.

For all preference assays (Figs. 1b, 1d, 1e, 2b), a male and two female flies were placed into a 38-mm diameter, 3-mm height circular chamber with sloping walls (courtship arena)[38]. The experimenter, who was blinded to the genotype of the flies, kept track of the females during the assay either by noting which female was introduced first to the courtship arena or by painting a small white dot on the thorax of the female 16–20 h before the start of the experiment under ice anaesthesia. Results were not affected by the method used to differentiate between females and the experimenter was unblinded only after analysis. The preference index is the amount of time the male spent courting one female subtracted from the amount of time spent courting the other female divided by the total time spent courting within a 10-min assay. The wide spread of the data reflects the fact that individual males will sometimes continue to pursue a single female throughout the assay even if both females are equivalent. In Extended Data Fig. 1 we plot the courtship indices (time spent courting/total time of assay) underlying the preference indices of Fig. 1.

For the tarsi ablation assays (Fig. 1b, Extended Data Fig. 1a–c), males were ice-anaesthetized 16–20 h before the start of the experiment and had either the distal three tarsal segments on both forelegs removed or a sham treatment that left their appendages intact. For single-choice assays (Extended Data Fig. 1b, c), the rear leg tarsi were ablated as a control. Males were then returned to a food vial to recover in isolation. Males without foreleg or rear leg tarsi still vigorously court females.

In *D. simulans* single-pair courtship assays (Figs. 1c, 3e, 4b, 4c; Extended Data Figs. 1b, 1c, 1e, 4e, 4f, 6f), a single virgin female and a *D. simulans* male were loaded into a courtship arena. Courtship index (time spent courting/total time together) was measured for the 10 min after the male was introduced into the chamber.

For the chaining assay (Extended Data Fig. 2i), eight males were loaded into a courtship arena and the chaining index (time in which at least three of the males were simultaneously courting each other divided by the total time of the assay) was measured for 10 min after the males were introduced into the chamber.

For the preference assays with perfumed females (Fig. 1e), we provided a male the choice between a *D. simulans* virgin female perfumed with 7,11-HD (7(*Z*),11(*Z*)-heptacosadiene, 10 mg ml^{-1}, Cayman Chemicals, 100462-58-6) or ethanol using a previously published protocol[8]. In brief, 7,11-HD or ethanol was added to 1 ml of ethanol in a 2-ml glass vial (Thermo Scientific, 03-377D), placed on ice and dried using a stream of nitrogen gas. Seven ice-anaesthetized female flies were placed in each vial, and the vials were then gently vortexed three times for 30 s before placing the females on food for an hour to recover. After perfuming, separate aspirators were used to handle the flies to avoid pheromone contamination. Courtship indices underlying courtship preference in Fig. 1e were plotted in Extended Data Fig. 1g.

For fru^{Gal4} optogenetic stimulation experiments (Fig. 2c), fru^{Gal4}>*UAS-CsChrimson.tdTomato* or fru^{Gal4} parental controls were reared in the dark for 3–7 days after eclosion. Male flies were transferred to food containing 400 μM all-*trans*-retinal (Sigma, R2500-10MG) 16–20 h before the assays[24]. Single male flies were loaded into a courtship arena and allowed to acclimate for 1 min. Flies were subsequently recorded for 7 min, alternating between 1 min of dim white light followed by 1 min with constant LED stimulation (530 nm Precision LED Spotlight with Uniform Illumination–PLS-0530-030-S, Mightex Systems at an intensity of 0.02 mW mm^{-2}). Genotypes were established using PCR screening of the UAS transgene. We quantified a courtship behaviour index, which represented the fraction of time a male spent performing courtship behaviours.

For *ppk23-Gal4* (Fig. 3e) and *R25E04-Gal4* (Extended Data Fig. 6f) optogenetic stimulation experiments, we used *D. simulans* w^+ *R25E04-Gal4*, *ppk23-Gal4* and *UAS-CsChrimson.tdTomato* parental stocks and *D. melanogaster* w^- *ppk23-Gal4* and *UAS-CsChrimson* parental stocks lacking balancer chromosomes. The original *D. simulans R25E04-Gal4*, *ppk23-Gal4* and *UAS-CsChrimson.tdTomato* parental stocks were in a background mutant for *white* (w^-), which exhibited extremely low courtship indices (around 5% on average), presumably owing to their low visual acuity. By contrast, *D. melanogaster* transgenic lines maintained robust courtship even in a *white* mutant background (data not shown). Therefore, we backcrossed *D. simulans* stocks to wild-type flies to generate w^+ strains and confirmed their genotype by PCR. All crosses were reared in the dark. Virgin male progeny were reared in isolation in the dark for 3–7 days after eclosion and then transferred to food containing 400 μM all-*trans*-retinal 16–20 h before the assays[24]. We found that *D. simulans* courtship was less robust under single-wavelength LED illumination or dim white-light illumination, so we conducted our assays using the same lighting conditions used for the non-optogenetic courtship assays (Logan Electric Slim Edge-Light Pad A-5A, 5400K, 6 klx). Single male flies were loaded into a courtship arena that contained a conspecific virgin female and the courtship index was calculated over a 10-min period after the male was introduced. Owing to the lack of balancers and visual markers in *D. simulans*, stable, homozygous stocks of most transgenes were difficult to generate. Therefore, progeny of crosses were a mix of wild-type, parental controls and experimental flies. The experimenter was blinded to the genotype of the flies until after the experiment. Genotypes were established using PCR amplification of the Gal4 and UAS transgenes. Males of all genotypes exhibited similar levels of locomotion when they were not courting (data not shown).

For optogenetic stimulation of P1 neurons in *D. simulans* (Fig. 4b, c, Extended Data Fig. 4c–f), we used *R71G01-Gal4>UAS-CsChrimson.mVenus* males that carried a wild-type (w^+) X chromosome. As in other behavioural experiments, P1 neuron-elicited courtship pursuit was far weaker in males mutant for *white* (data not shown). For optogenetic stimulation of P1 neurons in *D. melanogaster*, w^-; *R71G01-Gal4>UAS-CsChrimson.mVenus* males lacking balancer chromosomes were used. We observed a high degree of lethality in both the *D. melanogaster* and *D. simulans R71G01-Gal4>UAS-CsChrimson* crosses grown on standard fly food containing cornmeal (presumably owing to the low levels of retinal metabolized from vitamin A). Therefore, we grew these crosses on sugar-yeast food in the dark (per 1 l of water: 100 g Brewer's yeast, 50 g sucrose, 15 g agar, 3 ml propionic acid, 3 g *p*-hydroxy-benzoic acid methyl ester). Progeny of parental crosses were group-housed in the dark for 3–7 days after eclosion, before males were transferred to food containing 400 μM all-*trans*-retinal 48 h before the assays[24]. Single male flies were loaded into a courtship arena that contained either a virgin *D. simulans* female, virgin *D. melanogaster* female or a magnet (radius, 1 mm; height, 1 mm) rotating in a circle at 9 mm s^{-1} (ref. [25]). Upon loading the male fly into the chamber with the target, we alternated between 2 min of dim light (10 lx) and 2 min of bright light (6 klx) in a 14-min assay. Dim light was used because it was sufficient to enable males to visually track a target object but insufficient to optogentically activate the P1 neurons, as evidenced by the lack of courtship towards a magnet or *D. melanogaster* female before bright illumination. Assays were filmed (Sony Alpha 6) and later scored for courtship behaviour, binned in 1-s intervals. We calculated 'fraction courting' as a function of time by dividing the number of males courting during a one-second interval (aligned from the start of the assay) by the total males tested. Courtship indices were also calculated for each individual at

© 2018 Springer Nature Limited. All rights reserved.

different times relative to the optogenetic stimulation: 'pre' represents the courtship index of the 2 min before the first bright-light stimulus, 'stimulus' represents an average of the courtship indices during bright-light illumination period and 'post' represents an average of the courtship index after the bright-light illumination. For the parental controls (Extended Data Fig. 4d–f), we used w^+; *UAS-CsChrimson.mVenus* males grown in an identical manner to the experimental flies and similarly placed on retinal for 48 h. For the non-retinal controls (Extended Data Fig. 4d–f), w^+; *R71G01-Gal4>UAS-CsChrimson.mVenus* males were placed in a new vial of sugar-yeast food for 48 h before the experiment. To characterize evoked courtship as a function of light intensity (Fig. 4c), each experiment was initiated by illuminating for 2 min with dim light (10 lx) to establish a baseline, then at increasing intensities of 627-nm LED illumination with 2 min at each intensity, finally ending with 2 min of bright white-light illumination. A power meter (Coherent PowerMax-USD light sensor and Coherent PowerMax PC Software) was used to measure the intensity of 627-nm illumination in the behavioural chamber during the assay. To examine how elicited courtship depends on the speed of the magnet (Extended Data Fig. 4c), each male was given the opportunity to court a magnet moving at 0, 3, 6, 10 and 20 mm s^{-1} during bright white-light illumination. Magnet speed order was randomized and there were 1-min periods in between stimulus trials in which the light was off and the magnet was stationary.

Targeted mutagenesis and transformation in *D. simulans*. The protocols described below combine methods for CRISPR mutagenesis[39–41]. Supplementary Table 2 provides details of single-guide RNA (sgRNA) sequences, sgRNA primers and sequencing primers.

CRISPR guide RNAs had an 18–20 nucleotide target sequence and were flanked by a 3′ PAM sequence ('NGG') and a 5′ T7 RNA polymerase recognition sequence ('GG'). Before designing sgRNAs, Sanger sequencing was carried out across target genomic sites to identify single-nucleotide polymorphisms. Guide RNA template was amplified using KOD HotStart (Millipore, 71086-3) and 0.5 μM forward and reverse primers. Reactions were cycled on an Eppendorf MasterCycler (98 °C 30 s, 35 cycles of (98 °C 10 s, 60 °C 30 s, 72 °C 15 s), 72 °C 10 min, 4 °C hold) and then purified (PCR purification kit, Qiagen). In vitro transcription of 300 ng of sgRNA template DNA using T7 MEGAscript kit (Ambion) was carried out at 37 °C for 16–20 h. Turbo DNase was added for an additional 15 min at 37 °C before adding a 10% ammonium acetate stop solution. The RNA was phenol/chloroform extracted and precipitated with isopropanol at −20 °C for 16–20 h. Precipitated RNA was washed with 70% ethanol, resuspended with RNase-free water, and frozen in small aliquots at −80 °C for long-term storage. Before injection, the sgRNA was thawed on ice and purified using sodium acetate and ethanol before being resuspended in RNase free water.

CRISPR injection mixtures contained 300 ng μl^{-1} recombinant Cas9 protein (CP01, PNA Bio), 40 ng μl^{-1} sgRNA (per guide) and 125 ng μl^{-1} single-stranded DNA oligonucleotide. The CRISPR injection mixture was combined on ice and placed at −80 to −20 °C until the injection. φC-31 mediated recombination injection mixtures contained donor plasmid (1 μg μl^{-1}) and helper plasmid (1 μg μl^{-1}), both of which were purified using endotoxin-free plasmid prep kits (Qiagen). Rainbow Transgenic Flies, Inc. performed all injections.

To generate mutant alleles of *ppk23* and *Gr32a* (Extended Data Fig. 1d), we designed sgRNAs targeting three regions spanning 200 bp of the first exon for each gene. These sgRNAs were combined into a single cocktail and injected into approximately 200 wild-type *D. simulans* embryos. Only CRISPR guide sequences that generated the mutations are listed in Supplementary Table 2. The adult G_0 flies were individually crossed to wild-type male or virgin female flies. For each G_0 cross, we PCR screened 8–16 progeny (F_1s) for the presence of an insertion or deletion. Genomic DNA was extracted from the F_1 flies by removing a midleg, hindleg or wing and placing it into a well of a 96-well plate containing 20 μl of lysis buffer (10 mM Tris pH 8.2, 1 mM EDTA pH 8.0, 25 mM NaCl, 400 μg ml^{-1} Proteinase K). The fly was then placed in the corresponding well of a 96-well deep-well plate (Brandtech VWR, 80087-070) filled halfway with fly food and capped with cotton. The 96-well plate containing lysis buffer and fly legs or wings was then heated at 37 °C for 1 h followed by a 2-min heat inactivation at 95 °C. Genomic DNA (3.2 μl) from the reaction was used as the PCR template for a 25-μl reaction of Apex Taq Red Master Mix (Genesee Scientific, 42-138) for 35 cycles. The PCR screening primers spanned an approximately 400-bp region encompassing the three sgRNA target sites. To maximize the resolution of heterozygous indels, we ran the entire PCR reaction on a 2% agarose gel at 70 V. Using these specifications, the smallest indel we detected was around 20 bp. We backcrossed any flies that had a heterozygous mutation to wild-type flies and then homozygosed their progeny. Flies were Sanger-sequenced to determine if an in-frame stop codon was introduced. Homozygous stocks were genotyped and Sanger-sequenced for three generations to ensure that the population was pure.

For recombination into the *fru* locus, we prescreened sgRNAs to identify those that mediate efficient cutting. Nine sgRNAs were designed, six of which targeted the region upstream of the first coding exon and three of which targeted the first exon. Pools of three sgRNAs were injected into 100 embryos and genomic DNA was extracted from surviving flies. We first used the T7 endonuclease1 (T7E1) assay for preliminary qualitative analysis of cutting propensity (http://www.crispr-flydesign.org/t7-endo-i-assay/). Two positive hits from the T7E1 assays were analysed using MiSeq analysis[39], which revealed that over 95% of the reads in the PCR product were mutated. We used only these two sgRNAs (listed in Supplementary Table 2), one targeted to the first exon and one targeted upstream of the start codon of the first intron, for generating mutant flies.

To generate fru^{attP} flies, we integrated in a 200-bp single-stranded oligonucleotide designed to have the minimal 51 bp *attP* sequence[42], including a diagnostic restriction enzyme site and around 70 bp homology arms that flanked the CRISPR targeted site into the *fru* locus (Extended Data Fig. 2a). To generate $fru^{-/-}$ flies, we integrated a similar *attP*-containing oligonucleotide into the first exon of the Fru^M coding sequence, replacing the ATGATG start site with TTGTTG (Extended Data Fig. 2h), as has been reported previously in *D. melanogaster*[17]. The sgRNA, *attP*-oligo and Cas9 protein were injected into around 200 embryos. G_0s were singly crossed to wild-type virgin flies. F_1s with successful integration of the *attP* site were identified by PCR genotyping, isolated and sequenced using methods described above. fru^{attP} and $fru^{-/-}$ F_1s were backcrossed to wild-type flies and then homozygosed. Homozygous stocks were genotyped for three generations to ensure that the population was pure.

We used φC31-mediated recombination to integrate attB plasmids containing transgenes into the fru^{attP} locus (Extended Data Fig. 2a). We chose not to use eye colour as a visual marker to avoid complications of the *white* mutation on behaviour. To determine whether the transgene was homozygous, we screened F_1s using the protocol described above for the binary presence of a PCR product using one primer pair that spanned the transgene and the genomic locus. To create a stable stock of flies, we crossed homozygous virgin females to *D. simulans* males with a balancer allele on their 3rd chromosome (In(3R)Ubx, Flybase ID FBab0023784, UCSD Stock Center, 14021-0251.098). Progeny with the Ubx visible mutation were crossed together and subsequent progeny were genotyped.

attB-SAS-GFP (Extended Data Fig. 2a) was made by amplifying enhanced green fluorescent protein (eGFP) from pUAST-mCD8GFP using primers that attached a splice acceptor site[43] and Kozak sequence onto the 5′ end of the GFP and an SV40 termination sequence onto the 3′ end. Nested-PCR was performed to attach Gibson-assembly adaptors onto the GFP PCR product, which was then combined with PCR-linearized pHD-DsRed-attP using Gibson assembly (NEB). The plasmid was then digested with EcoRI and NotI to insert a 51-bp attB oligonucleotide with flanking EcoRI and NotI sites. The double-stranded oligonucleotide was made by annealing two single-stranded oligonucleotides together.

attB-SAS-Gal4 (Extended Data Fig. 2a) was made by cloning attB-SAS and Gal4 DNA fragments into pHD-DsRed cut with EcoRI and SpeI using Gibson Assembly (NEB). The digestion removed 3xP3-DsRed. The attB-SAS fragment was amplified from attB-SAS-GFP and the Gal4 fragment was amplified from pBPGUw.

We generated an *attP* landing site with an inactivated *EYFP* gene using CRISPR–Cas9 mutagenesis. We co-injected embryos of *D. simulans* strains carrying an *attP* landing site marked with *3XP3::EYFP* with p{CFD4-EYFP-3xP3::DsRed}[36] and *Cas9* mRNA and sibling-mated surviving adults. We screened for progeny with reduced or no *EYFP* expression in the eyes. Flies with $EYFP^-$ were bred to homozygosity and the *3XP3::EYFP* transgene in each line was re-sequenced to confirm the presence of the mutation and to confirm that the mutation did not disrupt the *attP* landing site.

To generate flies expressing GCaMP6s under UAS control, we co-injected p{GP-JFRC7-20XUAS-IVS-GCaMP6s}[44] and pBS130 (containing φC-31 integrase under control of a heat-shock promoter) into a *D. simulans attP*, *EYFP*- strain 2178 (ref. [36]) and screened for w^+ integrants. We generated one *D. simulans* UAS-CsChrimson transgenetic line by co-injecting p{20XUAS-IVS-CsChrimson.tdTomato}[45] and pBS130 into *attP*, $EYFP^-$ strain 2178 and screening for w^+ integrants. We generated a second *D. simulans* UAS-CsChrimson transgenetic line by co-injecting a piggyBac vector pBac(20xUAS-CsChrimson.mVenus, 3xp3::dsRed) and a piggyBac transposase helper plasmid into the Lethal Hybrid Rescue strain of *D. simulans*[46] and then screened for dsRed expression in the eyes. The CsChrimson.mVenus piggyBac insertion was mapped with TagMap[47] to base pair 23,569,712 on chromosome 3R.

The *D. simulans ppk23-Gal4* plasmids were generated by amplifying the 2.695-kb fragment upstream of the *D. simulans ppk23* promoter, analogous to previously published methods[12]. Using a BP-clonase Gateway reaction, the sim-*ppk23* promoter was recombined into pBPGUw (Addgene, 17575). φC31-mediated recombination was used to integrate sim into *D. simulans* attp2176[36] (Fig. 3d and Extended Data Fig. 3a–d), R25E04-Gal4, 3xp3::DsRed in *D. simulans* attP2176[36] (Extended Data Fig. 6e) and pBPGuW R71G01-Gal4 in *D. simulans* attP2176[36] (Extended Data Fig. 4a).

© 2018 Springer Nature Limited. All rights reserved.

Immunohistochemistry. To visualize *D. simulans fru*^{GFP} (Fig. 2a, Extended Data Fig. 2b–g), *D. melanogaster fru*^{Gal4}>*UAS-GCaMP* (Fig. 2a, Extended Data Fig. 2b–g), *D. simulans ppk23*^{Gal4}>*UAS-GCaMP* (Extended Data Fig. 3b) and *R25E04-Gal4*>*UAS-GCaMP* (Extended Data Fig. 6e), brains from 1–3-day-old adults were dissected in Schneider's medium for 1 h then immediately transferred to cold 1% paraformaldehyde (Electron Microscopy Sciences) and fixed for 16–20 h at 4 °C. Samples were then washed in PAT3 buffer (0.5% BSA, 0.5% Triton X-100, 1X PBS pH 7.4) three times, with the last two washes incubated for 1 h on a nutator at room temperature. Brains were blocked in 3% normal goat serum for 90 min at room temperature. Primary antibodies in 3% normal goat serum were incubated for 3 h at room temperature then left at 4 °C for 16–20 h. Primary antibodies used were 1:20 mouse anti-Brp (nc82, Developmental Studies Hybridoma Bank), 1:1,000 sheep anti-GFP (sim *fru*^{GFP}, mel *fru*^{Gal4}>*UAS-GCaMP*, *ppk23-Gal4*>*UAS-GCaMP* and *R25E04-Gal4*>*UAS-GCaMP*, Bio-Rad, 4745-1051) and 1:100 rabbit anti-GABA antibody (*D. simulans fru*^{GFP}, Sigma, A2052). Brains were then washed in PAT3 buffer. Samples were incubated in secondary antibody for 3 h at room temperature then for 5–7 days at 4 °C. Secondary antibodies used were 1:500 anti-sheep Alexa Fluor 488, anti-rabbit Alexa Fluor 546, anti-mouse Alexa Fluor 647 and anti-mouse Alexa Fluor 555 (ThermoFisher Scientific). Brains were washed in PAT3 buffer three times then once in 1X PBS, rotating at room temperature for 5 min. Samples were mounted in Vectashield (Vector Laboratories) in 5/8th inch hole reinforcements placed on glass slides. Images were captured on a Zeiss LSM 880 using a 40× objective.

Two-photon functional imaging. All imaging experiments were performed on an Ultima two-photon laser scanning microscope (Bruker Nanosystems) equipped with galvanometers driving a Chameleon Ultra II Ti:Sapphire laser. Emitted fluorescence was detected with either photomultiplier-tube or GaAsP photodiode (Hamamatsu) detectors. Images for ex vivo experiments were acquired with an Olympus 60×, 1.0 numerical aperture objective and in vivo experiments were acquired with an Olympus 40×, 0.8 numerical aperture objective (LUMPLFLN). All images were collected using PrairieView Software (Version 5.4) at 512 pixel × 512 pixel resolution with a frame rate from 0.2–0.4 Hz when imaging a region of interest (ROI) and 0.7–0.8 Hz when imaging the entire field of view. Saline (108 mM NaCl, 5 mM KCl, 2 mM $CaCl_2$, 8.2 mM $MgCl_2$, 4 mM $NaHCO_3$, 1 mM NaH_2PO_4, 5 mM trehalose, 10 mM sucrose, 5 mM HEPES pH 7.5, osmolarity adjusted to 275 mOsm) was used to bathe the brain for all imaging experiments unless otherwise noted.

To prepare flies for in vivo imaging of Fru^+ and $Ppk23^+$ sensory afferents in the ventral nerve cord (VNC; Fig. 3a–d, Extended Data Fig. 3e), the wings and all legs except one foreleg were removed from a CO_2-anaesthetized male. The single-legged male was tethered to a piece of clear packing tape covering a hole in the bottom of the modified 35-mm Petri dish using a hair placed across his cervical connectives. The body was oriented such that the ventral side faced the inside of the dish. A rectangular hole, the length and width of the body of the male fly, was cut from the tape and the fly was positioned such that the ventral half of the body was placed above the plane of the tape. Great care was taken to ensure that the foreleg was extended so that the tibia and femur did not contact the thorax. Small dots of UV-curable glue were used to secure the eyes, part of the thorax and the tip of the abdomen to the tape. The dish was then filled with saline and the cuticle covering the first thoracic ganglion was gently removed, taking care to not damage the foreleg nerve. The preparation was positioned on the two-photon microscope and an ROI was centred on the most ventral portion of the first thoracic ganglion of the VNC corresponding to the intact foreleg (Fig. 3a). To prepare a stimulating female abdomen, a pin was attached to the dorsal thorax of virgin female *D. melanogaster* or *D. simulans* fly with her head, wings and legs removed so that the abdomen could make contact with the distal tarsal segments of the foreleg of the male fly. To guide stimulation with the fly abdomen, an 850 nm IR light was used to illuminate the chamber. The fly was imaged from the side using a Point Grey Firefly camera mounted with a 1×-at-94 mm Infinistix lens fitted with a shortpass IR filter (850 nm, optical density 4, Edmund Optics) to block 925 nm two-photon laser illumination and viewed using FlyCapture2 Software (2.12.3.2). After recording a 10-s baseline, the experimenter gently tapped the female abdomen onto the tarsi of the experimental fly once every 10 s for 6–8 bouts. Three replicates per preparation (total 18–24 tapping bouts) were conducted with *D. simulans* and *D. melanogaster* stimuli interleaved.

Images and quantification of $Ppk23^+$ soma in the foreleg of the male were completed using a Zeiss Axioplan 2 scope under Nomarski optics and widefield fluorescence at 40× or 63×. Images were acquired through a Zeiss AxioCam and the Axiovision software (Extended Data Fig. 3a). Somata were counted only in the first three tarsal segments of the foreleg.

We modified published methods[12,21] for in vivo imaging of $Ppk23^+$ soma in the foreleg (Extended Data Fig. 3f). Male *ppk23-Gal4*>*UAS-GCaMP* flies were isolated as virgins, aged 3–6 days, CO_2-anaesthetized, decapitated and immobilized by folding a piece of Parafilm over the body such that the first five tarsal segments extended out of the Parafilm. The immobilized fly was placed on a glass coverslip for imaging using a monochromatic camera (Point Grey Research, Flir Chameleon 3). Pheromone was presented as follows: 1 µl of 7,11-HD or ethanol was pipetted onto a paper wick (Hampton Research) that had been trimmed such that one constituent fibre was exposed at the tip. Using a micromanipulator, the wick was brought into contact with one chemosensory sensillum on the 3rd tarsal segment of the foreleg. GCaMP responses were visualized using a 50× air objective using 488-nm LED illumination on a bright-field microscope (Scientifica). $\Delta F/F$ values were calculated using ImageJ as the maximum signal in the 30 s after pheromone presentation in accordance with published methods[12,21]. Without more precise genetic tools in *D. simulans*, we defined soma A as the soma that responded more strongly to 7,11-HD presentation, in accordance with previous work[12]. To demonstrate that the response of soma A was specific to the pheromone, we also presented the ethanol vehicle in which ethanol alone was adsorbed to a wick. The range of our maximum $\Delta F/F$ values for 7,11-HD stimulation were consistent with previously published results in *D. melanogaster*[12,21,48].

To prepare flies for in vivo imaging of the LPC, P1 neurons and mAL neurons in the central brain using both *fru*^{Gal4} and *R25E04-Gal4* neural drivers (Figs. 4d–h, 5d, Extended Data Figs. 4g–j, 6g–h), CO_2-anaesthetized 4–7-day-old males were affixed to a customized, milled tethering plate[49] using UV-curable glue around their head and thorax. Glue was cured in short bursts to minimize exothermic damage to the preparation and flies whose legs touched the glue were discarded. The proboscis was glued to the head, carefully avoiding the antennae, to minimize movement of the brain during imaging. Flies were given an hour to recover and were used only if they displayed vigorous activity post-tethering. A small hole in the head was opened under external saline using sharp forceps. Muscle 16, obstructing trachea, and fat were removed. The imaging plate had magnets inside to enable facile positioning under the 40× objective in the two-photon microscope. Using a micromanipulator, a styrofoam ball[50] floating on an air stream was positioned under the fly to permit standing and walking. A *D. melanogaster* or *D. simulans* virgin female tethered to a pin (see above for tethering details) was positioned in front of the tethered male using a micromanipulator. To stimulate tapping events, the female was moved in front of the male fly, who then freely tapped on her abdomen with his foreleg tarsi. The male fly was imaged from the side (see above) to facilitate positioning the ball and the abdomen stimulus during the experiment. After 4 to 5 s of baseline recording, the stimulus fly was presented to the tethered male for 2–5 s, allowing for several taps before being withdrawn. This was repeated nine times for each fly stimulus with *D. melanogaster* and *D. simulans* stimuli interleaved. An ROI was centred on the LPC or on the fasciculated projections from P1 neuron cell bodies to the LPC. We vetted our ability to reproducibly identify the characteristic processes of P1 neurons by first imaging them using *R71G01-Gal4*>*UAS-GCaMP* in *D. melanogaster* males (Extended Data Fig. 4g), which showed robust and specific responses in all parts of the P1 neurons to the taste of a *D. melanogaster* female. When imaging the fasciculated projections of P1 neurons, our field of view contained both the LPC and the P1 projections, enabling us to simultaneously record responses in the LPC and P1 neurons when the male tapped a female (Fig. 4g, h). We attempted to use the *R71G01-Gal4* driver for functional imaging of the P1 neurons in *D. simulans*, but observed no response to the taste of either a *D. melanogaster* or *D. simulans* female. Although this is consistent with the lack of pheromone-mediated excitation that we observed when imaging all Fru^+ neurons in the LPC or Fru^+ P1 neurons in *D. simulans*, we could not rule out that the lack of responses was due to weak expression of GCaMP. Notably, we observed pheromone responses using driver lines with similar levels of expression, such as *R25E04-Gal4*, in both *D. melanogaster* and *D. simulans*.

For experiments in which picrotoxin was used to block inhibition (Extended Data Fig. 6h), in vivo responses were recorded in the LPC before and after iontophoresis of picrotoxin unilaterally into the LPC (1 mM in saline, 3–5 pulses, 100 ms at 20 V). Local injection of picrotoxin had no noticeable effect on the behaviour of the male fly or on baseline fluorescence of the LPC, in contrast to bath application of picrotoxin (10 µM and 100 µM), which caused seizures in the fly and a marked, fluctuating increase in baseline fluorescence of the LPC (data not shown). Iontophoresis of saline had no effect on pheromone-evoked responses in either species (data not shown). Picrotoxin iontophoresis was based on previously published methods[51,52]. We did not attempt picrotoxin iontophoresis with *D. simulans R71G01-Gal4* because we could not confidently identify the LPC owing to weak expression of GCaMP.

To prepare flies for in vivo imaging of vAB3 (Fig. 5c, Extended Data Fig. 6a–d), 2–5-day-old male flies were briefly anaesthetized using CO_2 and then tethered used a previously described preparation[53] in which the male was affixed to a piece of tape covering a hole in the bottom of a modified 35-mm Petri dish using human hair placed across the cervical connectives. A small strip of tape was placed over the proboscis of the fly and two pieces of putty were placed next to the thorax to prevent the legs from becoming stuck onto the tape. A small hole above the head was precisely cut into the tape and the head was secured using two small dots of

© 2018 Springer Nature Limited. All rights reserved.

UV-curable glue that bridged the eyes and the tape. The dish was filled with saline and the head capsule was opened by carefully tearing off the flap of cuticle covering the dorsal portion of the head and removing any obstructing trachea and fat. The dish was placed under the microscope and the axonal tract of vAB3 projecting from the SEZ to the LPC was identified. We vetted our ability to reproducibly identify the characteristic morphology of vAB3 by first imaging the vAB3 axonal tract using *AbdB-Gal4>UAS-GCaMP* in *D. melanogaster* males (Extended Data Fig. 6a, b). Baseline fluorescence was recorded for 4 s before a female abdomen was presented to the male for him to tap. Trials were repeated three times for each female species and then $\Delta F/F$ responses were averaged.

Ex vivo stimulation of vAB3 (Fig. 5e, f, Extended Data Fig. 7) was performed as previously described[25]. A Grass stimulator was used to iontophorese acetylcholine (10 V, 200 ms) through a fine glass electrode positioned on the axons of the Ppk23$^+$ sensory neurons in the VNC (Fig. 5e, f, Extended Data Fig. 7). The stimulating electrode was filled with 10 mM acetylcholine, 10 mM glutamate or saline and Texas-Red Dextran BSA to facilitate positioning the electrode in the Fru$^+$ neuropil. The local nature of the stimulation combined with the anatomically segregated sensory innervation of the Ppk23$^+$ sensory neurons in the VNC facilitated restricted and reproducible stimulation. To functionally visualize responsive neurons in the brain, we imaged a z-plane every 5 μm and combined these to build a volume of the anterior around 100 μm of the brain. For quantitative comparisons of vAB3-evoked activity in specific neural populations across individuals (Fig. 5f, Extended Data Fig. 7), single z-planes were recorded using a 40× objective at 2× zoom with an ROI of 300 × 300 pixels. Given that P1 soma and fasciculated processes reside on the posterior side of the brain, when imaging P1 and vAB3 neurons in response to vAB3 stimulation, we rotated the brain 180° around the cervical connectives.

For two-photon severing of mAL (Fig. 5f, Extended Data Fig. 7), the brain was pinned ventral side up and we focused 925-nm light on a small ROI encompassing only the mAL axon tract at 8× optical zoom. The mAL axon tract could be readily identified by its characteristic morphology. For two-photon severing of vAB3 or a mock Fru$^+$ neuron (Extended Data Fig. 7c), the VNC and brain were pinned ventral side up. We validated that vAB3 axons could be reproducibly identified within the VNC by performing initial experiments in *AbdB-Gal4>UAS-Tomato/fru^{LexA}>LexAOP-GCaMP D. melanogaster* males in which vAB3 neurons are anatomically marked by Tomato expression. We found that vAB3 axons were always robustly activated by acetylcholine iontophoresis and have a characteristic position within the ventral cord that enabled their identification even in the absence of an anatomic marker. We focused 925-nm light on a small ROI encompassing either the vAB3 axon tract or the tract of a Fru$^+$ neuron more lateral than vAB3. We then switched the laser to 850 nm and imaged using short (<1 s) pulses until a cavitation bubble was observed. After switching back to 925 nm and zooming out, if the axon tract was successfully severed, we observed a marked increase in baseline fluorescence due to Ca^{2+} influx into the neurons and activating GCaMP. Because vAB3 neurons project bilaterally, we also severed the corresponding axon tract on the contralateral side of the brain. To image P1 neurons after severing mAL, we re-pinned the brain such that the dorsal side of the brain and the ventral side of the VNC were facing up, rotating around the cervical connectives, inserted the stimulating electrode in the VNC and recorded activity in P1 neurons and vAB3 neurons. vAB3 activation was not affected by mAL severing (data not shown).

Dye-filling of neural tracts using Texas-Red Dextran (100 mg ml^{-1}, Invitrogen) was performed as previously described[54]. For dye-filling specific neural populations, we targeted the fasciculated bundle of P1 neurons projecting from the somata (Fig. 4a, h, Extended Data Fig. 5a), the segregated vAB3 terminals in the VNC (Fig. 5b, Extended Data Fig. 5b) and the characteristic mAL axonal bundle projecting between the SEZ and LPC (Fig. 5b, Extended Data Fig. 5c) with a dye-filled electrode. To photolabel neurons, we expressed sPA-GFP in Fru$^+$ neurons, located the neural structure of interest using 925-nm laser illumination, a wavelength that does not cause substantial photoconversion, defined an ROI in PrairieView software in a single z-plane, and exposed the target area to 710-nm light (around 10–30 mW at the back aperture of the objective) 100–300 times over the course of several rounds of photoactivation. We allowed diffusion of the photoconverted fluorophores throughout the targeted neurons for 30–60 min and then imaged at 925 nm. All anatomical images are maximum projections of z-stacks with 1-μm steps. Autofluorescence from the glial sheath and basal fluorescence from non-dye-filled structures were masked for clarity.

Unless stated, anatomical images were acquired on the 2P microscope using standard techniques.

Imaging analysis. To analyse functional responses, we calculated $\Delta F/F$ for each frame of calcium-imaging time-courses, using the second to sixth frames as the baseline. The peak $\Delta F/F$ value was defined as the maximum value within the time during which the stimulus was presented, which could vary due to inconsistent tapping rate of a male fly. Heat maps (ΔF) were generated using the maximum projection of two frames of baseline subtracted from the maximum projection of the two frames encompassing the peak fluorescence in response to a stimulus. Arbitrary units (AU) correspond to 1/100th of the minimum displayed value and maximum displayed value in the display range set in FIJI.

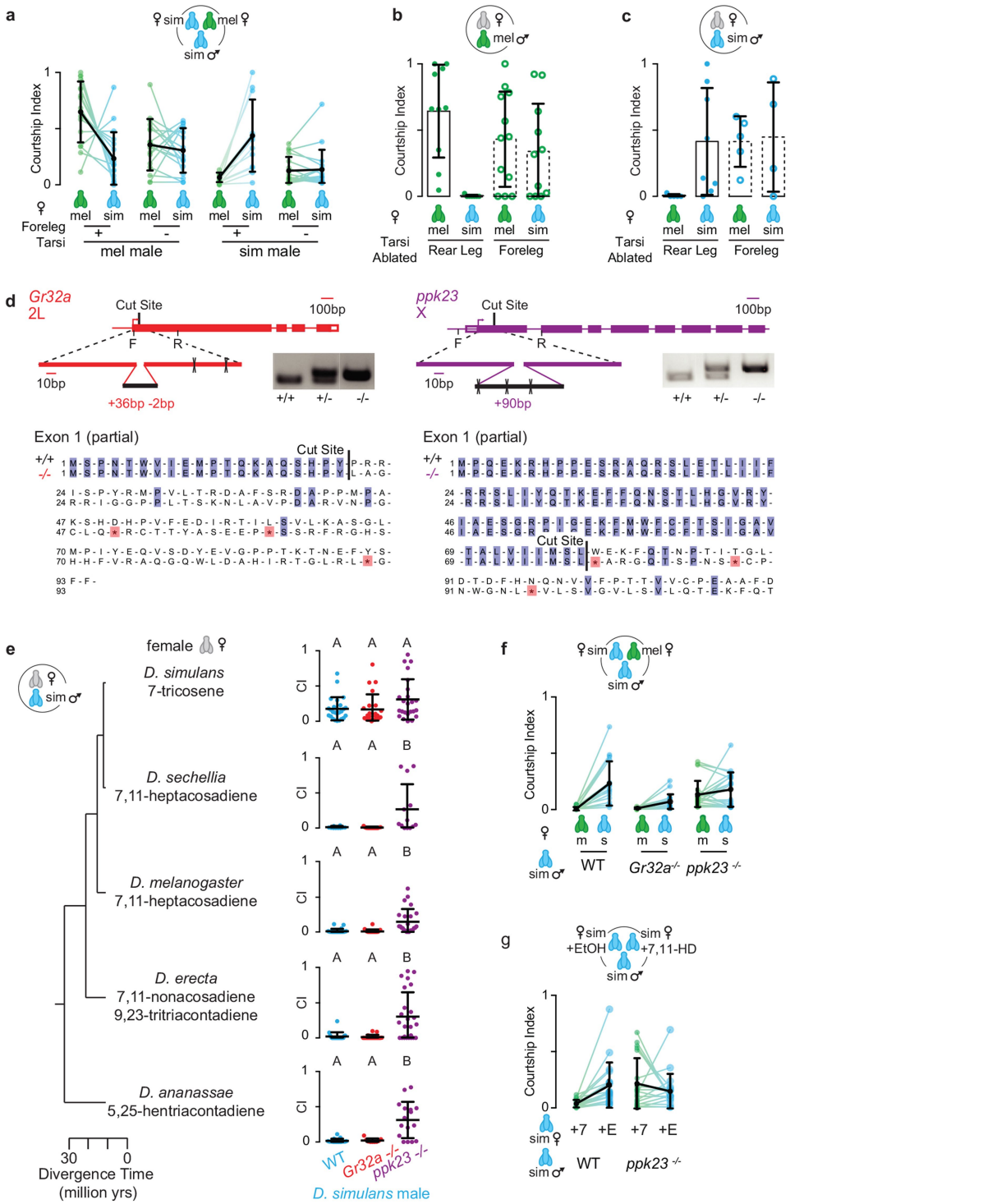
Statistics and reproducibility. For courtship indices, each data point represents a biological replicate of an individual male's courtship behaviour. For in vivo and ex vivo functional assays (Figs. 3b–d, 4e, f, 5c, d, f; Extended Data Figs. 3f, 6b, c, h, 7c, e), each pair of dots represents the average response for a given preparation (biological replicate). In Fig. 4g, h and Extended Data Figs. 3e, 4h–j, 6d, g, 7c–g, individual stimulations were plotted. Each representative example was replicated more than five times in independent experiments (Figs. 2a, 3a–c, 4a, e–h, 5b, e; Extended Data Figs. 2b–g, 3a–d, 4a, g, 5, 6a, e, 7a–c and Supplementary Videos 1–5). Sample sizes were based on pilot experiments. For all behavioural experiments $n > 15$, and for all in vivo imaging experiments $n > 5$. We used the PRISM software package to graph and statistically analyse data. Before statistical analysis, we tested whether the values were normally distributed using D'Agostion–Pearson omnibus and Shapiro–Wilk normality tests. When data were normally distributed, we used parametric tests; when data were not normally distributed, we used non-parametric tests. We adjusted P values accordingly when multiple comparisons were conducted. See Supplementary Table 1 for more details on statistical tests and P values.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. Datasets generated during the current study are available from the corresponding author on reasonable request.

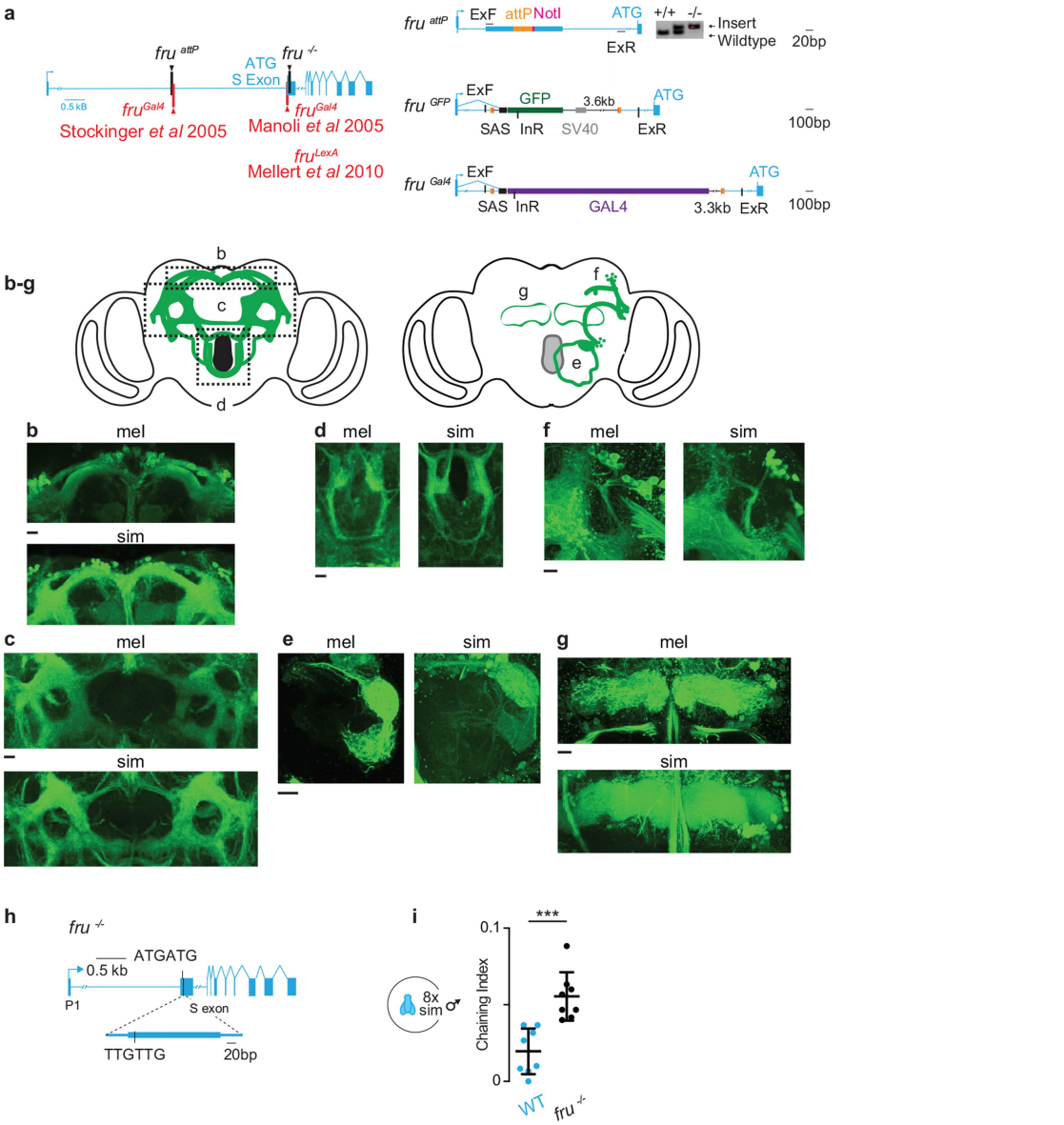
36. Stern, D. L. et al. Genetic and transgenic reagents for *Drosophila simulans*. *G3* **7**, 1339–1347 (2017).
37. Mellert, D. J., Knapp, J.-M., Manoli, D. S., Meissner, G. W. & Baker, B. S. Midline crossing by gustatory receptor neuron axons is regulated by *fruitless*, *doublesex* and the Roundabout receptors. *Development* **137**, 323–332 (2010).
38. Agrawal, S., Safarik, S. & Dickinson, M. The relative roles of vision and chemosensation in mate recognition of *Drosophila melanogaster*. *J. Exp. Biol.* **217**, 2796–2805 (2014).
39. Kistler, K. E., Vosshall, L. B. & Matthews, B. J. Genome engineering with CRISPR–Cas9 in the mosquito *Aedes aegypti*. *Cell Rep.* **11**, 51–60 (2015).
40. Gratz, S. J. et al. Highly specific and efficient CRISPR/Cas9-catalyzed homology-directed repair in *Drosophila*. *Genetics* **196**, 961–971 (2014).
41. Bassett, A. R., Tibbit, C., Ponting, C. P. & Liu, J. L. Highly efficient targeted mutagenesis of *Drosophila* with the CRISPR/Cas9 system. *Cell Rep.* **4**, 220–228 (2013).
42. Thorpe, H. M., Wilson, S. E. & Smith, M. C. M. Control of directionality in the site-specific recombination system of the *Streptomyces* phage φC31. *Mol. Microbiol.* **38**, 232–241 (2000).
43. Venken, K. J. T. & Bellen, H. J. Transgenesis upgrades for *Drosophila melanogaster*. *Development* **134**, 3571–3584 (2007).
44. Chen, T.-W. et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
45. Klapoetke, N. C. et al. Independent optical excitation of distinct neural populations. *Nat. Methods* **11**, 338–346 (2014).
46. Brideau, N. J., Flores, H. A., Wang, J., Maheshwari, S., Wang, X. & Barbash, D. A. Two Dobzhansky–Muller genes interact to cause hybrid lethality in Drosophila. *Science* **314**, 1292–1295 (2006).
47. Stern, D. L. Tagmentation-based mapping (TagMap) of mobile DNA genomic insertion sites. Preprint at https://doi.org/10.1101/037762 (2017).
48. Vijayan, V., Thistle, R., Liu, T., Starostina, E. & Pikielny, C. W. *Drosophila* pheromone-sensing neurons expressing the *ppk25* ion channel subunit stimulate male courtship and female receptivity. *PLoS Genet.* **10**, e1004238 (2014).
49. Maimon, G., Straw, A. D. & Dickinson, M. H. Active flight increases the gain of visual motion processing in *Drosophila*. *Nat. Neurosci.* **13**, 393–399 (2010).
50. Seelig, J. D. et al. Two-photon calcium imaging from head-fixed *Drosophila* during optomotor walking behavior. *Nat. Methods* **7**, 535–540 (2010).
51. Hill, R. G., Simmonds, M. A. & Straughan, D. W. Antagonism of GABA by picrotoxin in the feline cerebral cortex. *Br. J. Pharmacol.* **44**, 807–809 (1972).
52. Crossman, A. R., Walker, R. J. & Woodruff, G. N. Picrotoxin antagonism of γ-aminobutyric acid inhibitory responses and synaptic inhibition in the rat substantia nigra. *Br. J. Pharmacol.* **49**, 696–698 (1973).
53. Cohn, R., Morantte, I. & Ruta, V. Coordinated and compartmentalized neuromodulation shapes sensory processing in *Drosphila*. *Cell* **165**, 715–729 (2015).
54. Ruta, V. et al. A dimorphic pheromone circuit in *Drosophila* from sensory input to descending output. *Nature* **468**, 686–690 (2010).

© 2018 Springer Nature Limited. All rights reserved.



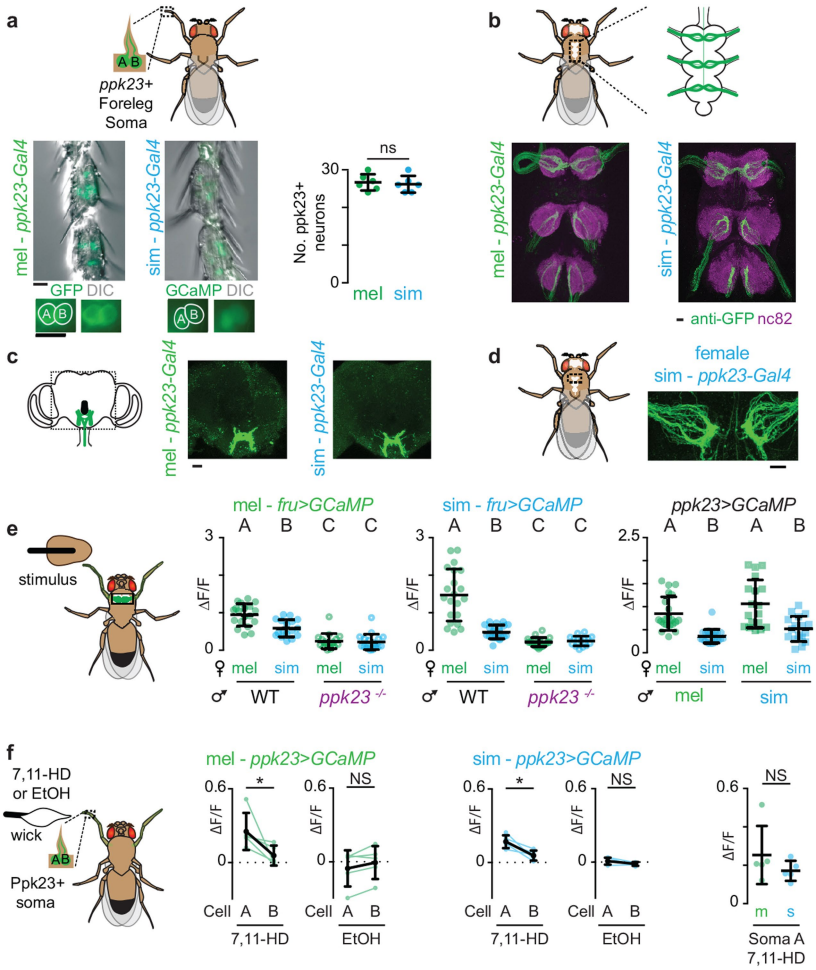
Extended Data Fig. 1 | Pheromone regulation of *D. simulans* courtship. Mutant males and males lacking foreleg tarsi still court, but display altered courtship preferences. **a**, Courtship indices of males with foreleg tarsi intact (+) or surgically removed (−). Data are replotted from Fig. 1b. **b**, **c**, Courtship indices of *D. melanogaster* (**b**) and *D. simulans* (**c**) males with either foreleg tarsi or rear leg tarsi ablated towards *D. melanogaster* or *D. simulans* females. **d**, Schematic of CRISPR–Cas9-induced mutations (top) in *Gr32a* (left) and *ppk23* (right) gene loci. Cas9 was targeted by gRNA to the first exon (cut site) of *Gr32a* or *ppk23* resulting in a 36-bp insertion/2-bp deletion in the *Gr32a* coding sequence and 90-bp insertion into the *ppk23* coding sequence. Both indels generated in-frame stop codons (bottom, asterisk highlighted red in resulting amino acid sequence). Forward (F) and reverse (R) genotyping primers are marked with a line. **e**, Courtship indices (CI) towards females of different *Drosophila* species by wild-type (WT), *Gr32a*$^{-/-}$ and *ppk23*$^{-/-}$ *D. simulans* males. The major pheromone carried by each female is listed. **f**, Courtship indices of *D. simulans* males towards *D. melanogaster* and *D. simulans* females in preference assays. Data are replotted from Fig. 1d. **g**, Courtship indices of *D. simulans* males towards *D. simulans* females perfumed with 7,11-HD (7, green) or ethanol (E, EtOH, blue). Data are replotted from Fig. 1e. The statistical test performed in **e** was a Kruskal–Wallis test, and different letters mark significant differences. Data are mean and s.d., with individual data points shown. Lines connect courtship indices of the same male towards the different female targets in a preference assay. Because the male can only court one female at a time, the paired points are inherently interdependent on each other, therefore inappropriate for statistical analysis. See Supplementary Table 1 for details of statistical analyses.

© 2018 Springer Nature Limited. All rights reserved.



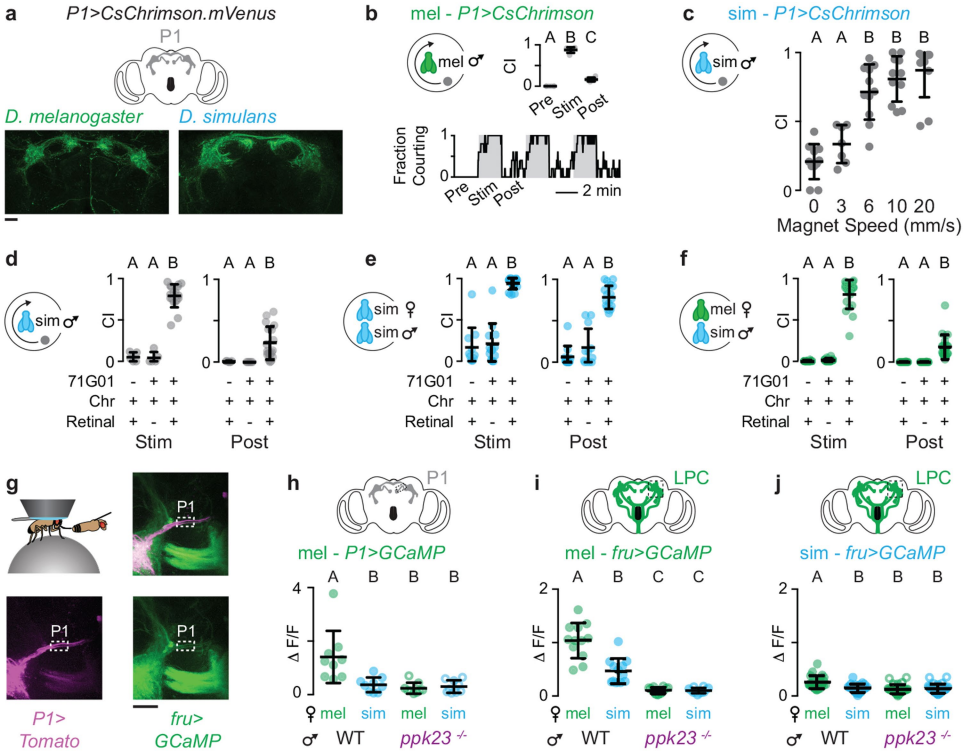
Extended Data Fig. 2 | Anatomic and functional conservation of Fru$^+$ neurons. a, Left, schematic of chromosomal location of *fru*^{attP} and *fru*$^{-/-}$ integration sites in *D. simulans* and previously generated *fru*^{Gal4} and *fru*^{LexA} transgenes in *D. melanogaster*; right, schematic of *attP* oligonucleotide integrated into the *fru* intron to generate *fru*^{attP} allele and subsequent integration of attB plasmids (right). ExF and ExR are primers located in the genome and InR is a primer located inside the transgene. **b**–**g**, Maximum intensity confocal (**b**–**d**) and two-photon stacks (**e**–**g**) of anatomically defined regions of Fru$^+$ neuropil in *D. melanogaster* *fru*^{Gal4}>*UAS-GCaMP* and *D. simulans fru*^{GFP} males: LPC (**b**, **c**), SEZ (**d**), antennal lobe (**e**), lateral horn (**f**) and mushroom body γ-lobes (**g**). Scale bars, 10 μm. **h**, *D. simulans fru*$^{-/-}$ was generated by integrating an oligonucleotide that deleted codons 1 and 2 of the first exon, introducing a frame-shift mutation. **i**, Male–male chaining indices of wild-type (+/+) and *fru*$^{-/-}$ males. A paired *t*-test was used, data are mean and s.d. and individual data points are shown. See Supplementary Table 1 for details of statistical analyses.

© 2018 Springer Nature Limited. All rights reserved.



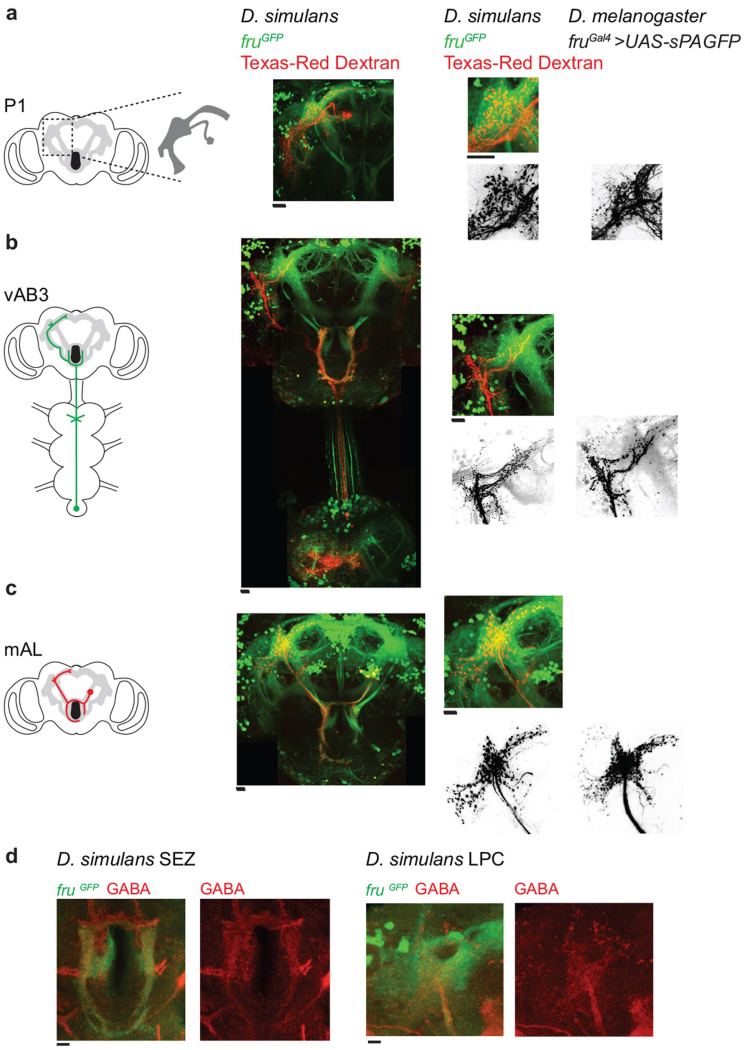
Extended Data Fig. 3 | Conserved anatomy and functional tuning of Ppk23^{+}Fruitless^{+} foreleg sensory neurons. a–c, *ppk23* promoter expression in *D. melanogaster* and *D. simulans* males in forelegs (**a**), ventral nerve cord (VNC, **b**) and brain (**c**). GFP, green; differential interference contrast, grey; neuropil counterstain, magenta. **a**, Forelegs (top left) with paired soma (bottom left) and quantification of number of Ppk23^{+} sensory neuron soma in the first three tarsal segments of the foreleg (right). **d**, Ppk23 neuron innervation in the first thoracic ganglion of the VNC of *D. simulans* females. Ppk23^{+} sensory neurons display a characteristic sexually dimorphic expression pattern in the VNC, where they do not cross the midline in females, but do in males. **e**, Schematic of VNC imaging preparation (left); functional responses evoked by individual taps of a female abdomen in Fru^{+} neurons (middle) and Ppk23^{+} neurons (far right) in the VNC of wild-type and *ppk23*$^{-/-}$ *D. melanogaster* and *D. simulans* males. Data replotted from Fig. 3b–d. **f**, Schematic of Ppk23^{+} somatic imaging preparation (left) and functional responses of the paired neurons (cell A and B, see Methods) within a sensory bristle stimulated with 7,11-HD or ethanol (middle); comparison of 7,11-HD responses in Ppk23^{+} soma across species (far right). Statistical tests used were an unpaired *t*-test (**a**), a Kruskal–Wallis test, for which different letters mark significant differences (**e**), and paired and unpaired *t*-tests (**f**). Scale bars, 10 μm. See Supplementary Table 1 for details of statistical analyses.

© 2018 Springer Nature Limited. All rights reserved.



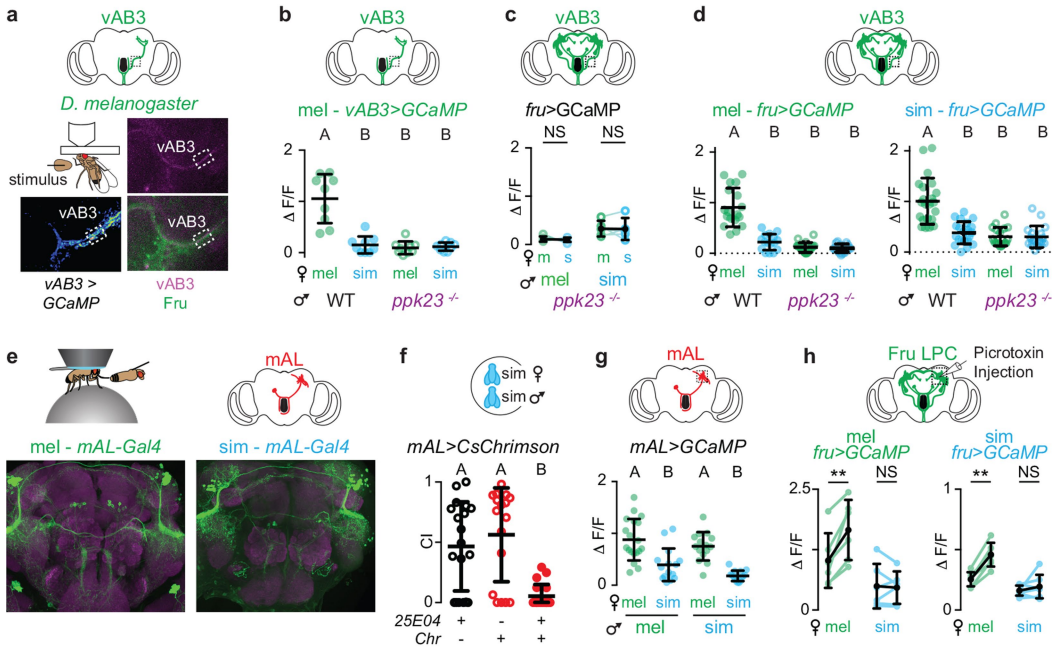
Extended Data Fig. 4 | Behavioural and functional analysis of P1 neurons. a–f, Anatomy (**a**) and optogenetic behavioural manipulations (**b–f**) of P1 neurons using *R71G01-Gal4* to drive the expression of CsChrimson. **b**, Courtship indices (top, right) towards a rotating magnet by *D. melanogaster* males pre-, during, and post-P1 neuron optogenetic stimulation. Fraction of male flies courting (bottom; grey boxes indicate bright light illumination, see Methods). **c**, Courtship indices towards a magnet moving at different speeds during optogenetic P1 neuron stimulation in *D. simulans* males. **d–f**, Comparison of courtship indices towards magnet (**d**), *D. simulans* female (**e**), or *D. melanogaster* female (**f**) by *D. simulans* males of denoted genotypes, either fed or not fed retinal. **g**, Experimental set-up used to measure pheromone responses in the P1 neurons in vivo (top left) and overlay of the Fru^{+} (green) neurons and fasciculated P1 neuron processes (magenta). The white boxes indicate the approximate ROI used to image P1 responses. **h–j**, Functional responses evoked by individual taps of a female abdomen in P1 neurons (**h**) and Fru^{+} neurons in the LPC (**i**, **j**) of wild-type and *ppk23*$^{-/-}$ *D.melanogaster* and *D. simulans* males (data replotted from Fig. 4e, f). Statistical tests used were the Kruskal–Wallis test (**b**, **d–f**, **h–j**) and a one-way ANOVA (**c**). Data are mean and s.d., with individual data points shown. See Supplementary Table 1 for details of statistical analyses.

© 2018 Springer Nature Limited. All rights reserved.



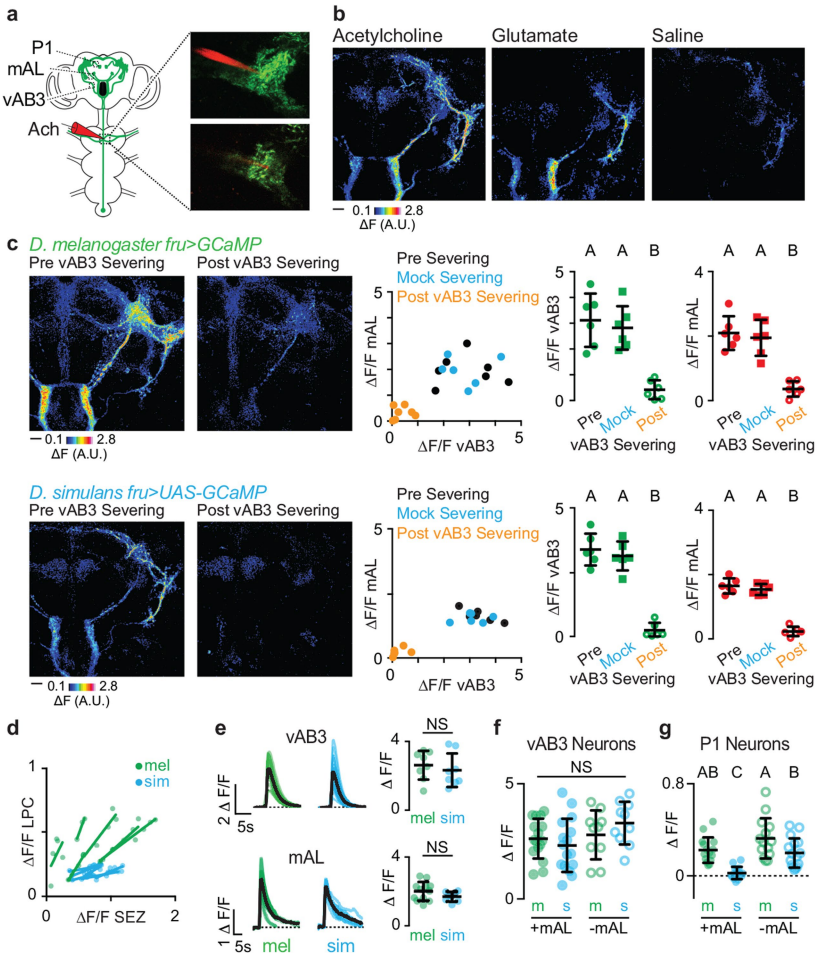
Extended Data Fig. 5 | Anatomy of P1, vAB3 and mAL neurons in *D. simulans* and *D. melanogaster* males. a–c, Detailed anatomic images of P1 neurons (**a**), vAB3 neurons (**b**) and mAL neurons (**c**). Schematic of neural anatomy (left), Texas-Red dextran dye-fill (red) in *D. simulans* fru^{GFP} (green) males (middle-left), magnified view of labelled neurons in the LPC showing dye-filled neurons (red) and Fru^{+} neurons (green) or just dye-filled neurons (black, middle-right) and photoactivated neurons in *D. melanogaster* LPC (black, right). **d**, Antibody staining of *D. simulans* Fru^{+} neurons (anti-GFP, green) with anti-GABA (red) in the SEZ and the LPC demonstrating that mAL neurons are GABAergic and thus inhibitory. Scale bars, 10 μm.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 6 | Pheromone responses in central neurons of *D. melanogaster* and *D. simulans* males. a, Schematic of in vivo preparation used to measure pheromone responses in vAB3 processes in the brain (top). Representative fluorescence increase of vAB3 responses in a *D. melanogaster* male evoked by tapping a *D. melanogaster* female (bottom left). GCaMP was expressed in vAB3 neurons using the AbdB-Gal4 driver. Anatomy of fasciculated vAB3 processes co-labelled by fru^{Gal4} (green) and *AbdB-Gal4* (magenta) in the same in vivo preparation used for imaging (right). The white box indicates the approximate ROI analysed for functional imaging. **b**–**d**, Functional responses evoked by the taste of female pheromones in the vAB3 processes of wild-type and $ppk23^{-/-}$ males. **b**, Functional responses evoked by individual taps in vAB3 neurons in *D. melanogaster* labelled using *AbdB-Gal4*. **c**, Average responses of vAB3 neurons in $ppk23^{-/-}$ *D. melanogaster* and *D. simulans* males in response to the taste of *D. melanogaster* (m) and *D. simulans* (s) females. GCaMP was expressed in vAB3 neurons using fru^{Gal4}. **d**, Functional responses evoked by individual taps in vAB3 neurons in wild-type and $ppk23^{-/-}$ mutant males. Data replotted from Fig. 5c and Extended Data Fig. 6c. **e**, Expression of *25E04-Gal4>UAS-GCaMP* (green) with neuropil counterstain (magenta) in the brains of *D. melanogaster* (left) and *D. simulans* (right) males. **f**, Courtship indices towards conspecific females during optogenetic activation of mAL neurons in *D. simulans* males with parental controls. **g**, Functional responses evoked by individual taps in mAL neurons. Data replotted from Fig. 5d. **h**, Average $\Delta F/F$ responses in Fru^{+} neurons of the LPC evoked by the taste of a female before and after local injection of picrotoxin, a $GABA_A$ receptor antagonist, into the LPC. In males of both species, application of picrotoxin increased responses only to *D. melanogaster* female stimuli. Lines connect average functional responses in the same male towards the different female targets. Statistical tests used were the Kruskal–Wallis test, in which different letters mark significant differences (**b**, **d**, **f**, **g**), and the paired *t*-test (**c**, **h**). Data are mean and s.d., with individual data points shown. See Supplementary Table 1 for details of statistical analyses.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 7 | Functional responses of Fru$^+$ neurons to direct vAB3 stimulation in *D. melanogaster* and *D. simulans* males. **a**, Schematic and representative image depicting direct stimulation of vAB3 neurons by iontophoresis of neurotransmitter onto their dendrites within the VNC. Maximum intensity *z*-projection (top) and single *z*-plane (bottom) showing electrode placement in the VNC. The electrode is filled with neurotransmitter and Texas-red dye (red) to enable precise targeting in the Fru$^+$ neuropil (green). **b**, Representative multi-plane fluorescence increase of Fru$^+$ brain neurons in *D. simulans* males when the VNC is stimulated with acetylcholine (left), glutamate (middle) and saline (right) iontophoresis. **c**, To test the necessity of vAB3 in propagating signals from the VNC to the higher brain, we compared response profiles in the brain before severing vAB3 (black), after severing a nearby Fru$^+$ axon (mock control, blue) and then after severing vAB3 axons (orange) in *D. melanogaster* males (top) and *D. simulans* males (bottom). Representative multi-plane fluorescence increase of Fru$^+$ neurons (left) shows the loss of evoked functional responses in the brain after severing vAB3. The graph (middle) depicts the relationship between average $\Delta F/F$ responses in vAB3 and mAL. Average $\Delta F/F$ responses of vAB3 (green) and mAL (red) neurons evoked by vAB3 stimulation, before and after severing vAB3 (right). Responses were lost in both neural populations in both species after vAB3 was severed but not in the mock control. **d**, The relationship of functional responses in the SEZ and LPC of *D. melanogaster* and *D. simulans* males evoked by direct vAB3 stimulation. The dots on the graph represent different stimulation intensities and the trend lines relate the responses of individual *D. melanogaster* (green) and *D. simulans* (blue) males. **e**, Left, the response of vAB3 (top) and mAL (bottom) axonal tracts in response to vAB3 stimulation in *D. melanogaster* (green) and *D. simulans* (blue) males. Coloured lines represent single stimulations and black lines represent the average. Right, peak $\Delta F/F$ plotted with coloured dots representing the average response per fly and data are mean and s.d. Individual data points are shown. **f**, **g**, Comparison of vAB3-evoked responses in *D. melanogaster* and *D. simulans* males in vAB3 neurons (**f**) and P1 neurons (**g**) before (+) and after (−) mAL severing. Statistical tests used were the Kruskal–Wallis test, in which different letters mark significant differences (**c**, **f**, **g**), and an unpaired *t*-test (**e**). Scale bars, 10 μm. See Supplementary Table 1 for details of statistical analyses.

© 2018 Springer Nature Limited. All rights reserved.

ARTICLE

https://doi.org/10.1038/s41586-018-0333-6

Cryo-EM structure of a fungal mitochondrial calcium uniporter

Nam X. Nguyen[1,2], Jean-Paul Armache[3], Changkeun Lee[1,2,6], Yi Yang[1,2], Weizhong Zeng[1,2], Vamsi K. Mootha[4], Yifan Cheng[3], Xiao-chen Bai[2,5]* & Youxing Jiang[1,2]*

The mitochondrial calcium uniporter (MCU) is a highly selective calcium channel localized to the inner mitochondrial membrane. Here, we describe the structure of an MCU orthologue from the fungus *Neosartorya fischeri* (NfMCU) determined to 3.8 Å resolution by phase-plate cryo-electron microscopy. The channel is a homotetramer with two-fold symmetry in its amino-terminal domain (NTD) that adopts a similar structure to that of human MCU. The NTD assembles as a dimer of dimers to form a tetrameric ring that connects to the transmembrane domain through an elongated coiled-coil domain. The ion-conducting pore domain maintains four-fold symmetry, with the selectivity filter positioned at the start of the pore-forming TM2 helix. The aspartate and glutamate sidechains of the conserved DIME motif are oriented towards the central axis and separated by one helical turn. The structure of NfMCU offers insights into channel assembly, selective calcium permeation, and inhibitor binding.

Mitochondria can take up large amounts of calcium from their environment, which can modulate ATP production, alter cytoplasmic Ca^{2+} dynamics, and trigger cell death[1,2]. Calcium enters the mitochondrial matrix through the mitochondrial calcium uniporter, a highly selective calcium channel that is localized to the inner mitochondrial membrane[3]. In humans, the uniporter is a protein complex, or 'uniplex', consisting of at least four components: the ion conducting pore (MCU)[4,5], an essential membrane spanning subunit (EMRE)[6], and peripheral membrane gate-keeping proteins (MICU1 and MICU2)[7,8]. Whereas MCU is found in all major eukaryotic taxa, EMRE is metazoan-specific and is required for the conductivity of MCU in these organisms[6,9]. By contrast, the MCU homologue in *Dictyostelium*, whose genome lacks EMRE, is sufficient to operate as a pore-forming channel[10]. Uniporter activity has diverged in fungi, even being lost entirely in certain lineages such as *Saccharomyces cerevisiae*[11]. In most fungi, genome sequence analysis indicates that the pore-forming MCU is the only component of the uniporter[12] and is likely to represent the minimal channel component of the uniporter for calcium uptake. A recent study used functional mutagenesis to demonstrate that the MCU orthologue from the fungus *Aspergillus fumigatus* can mediate Ca^{2+} transport into mitochondria[13].

Notably, MCU exhibits no discernible sequence homology to other cation channels[12], making it difficult to predict its structure and function using computational tools. Although the structures of various isolated MCU domains or the auxiliary component of the uniplex have been reported—the NTD of human MCU[14,15], the pore domain of an N-terminal deletion of the *C. elegans* MCU (cMCU-ΔNTD)[16], and human MICU1[17]—fundamental questions about the channel's assembly, gating, and ion permeation remain unanswered. The recent NMR structure of cMCU-ΔNTD revealed a pentameric channel pore with an occluded ion pathway, and EMRE was suggested to be the key to driving the pore from being in the occluded to an open, ion-conductive state[16]; however, there are no structural data on metazoan MCU and EMRE together. Mounting evidence for the pathophysiological relevance of the uniporter complex[1,18] has provided further impetus for determination of the structure of the intact channel. To address these longstanding questions, we determined the cryo-EM structure of an MCU homologue from *N. fischeri* (NfMCU) and also performed in-depth functional studies on this channel. The structure defines the architecture of the channel subunit of the MCU and provides insights into channel assembly and function.

Biochemistry and structure determination

The architectural and functional complexity of metazoan MCU led us to take a reductionist approach and to characterize MCU orthologues from fungi, whose genomes encode only the ion-conducting pore subunit[12], with the goal of understanding the basic principles that govern channel assembly and function. Among those tested, the 488-residue wild-type MCU orthologue from *N. fischeri*, NfMCU, showed the best biochemical properties in terms of protein expression and stability and was modified for structural studies (see Methods). Unlike the pentameric cMCU-ΔNTD from the NMR study[16], NfMCU purified as a tetramer in solution (Extended Data Fig. 1). Additionally, it conferred uniporter function to bacteria, and NfMCU proteoliposomes recapitulated the voltage-dependent Ca^{2+} uptake property of the MCU[19] (see Methods and Extended Data Fig. 2).

Determination of the structure of NfMCU using single particle cryo-EM was technically challenging for the following reasons: small protein size of 180 kDa for a channel tetramer; symmetry mismatch between the pore and soluble domains; and an elongated shape with minimal features. In addition, detergent-solubilized NfMCU aggregated when frozen and NfMCU in amphipols exhibited orientation bias. We first overcame the stability issue by reconstituting the channel into nanodiscs but, owing to the channel's flexibility in nanodiscs, whose size is much larger than the transmembrane part of the channel, we could only determine the structure to around 15 Å resolution. This led us to attempt reconstitution using the newly developed scaffolding protein saposin (see Methods)[20], which we rationalized could adapt its

[1]Howard Hughes Medical Institute and Department of Physiology, University of Texas Southwestern Medical Center, Dallas, TX, USA. [2]Department of Biophysics, University of Texas Southwestern Medical Center, Dallas, TX, USA. [3]Howard Hughes Medical Institute and Keck Advanced Microscopy Laboratory, Department of Biochemistry and Biophysics, University of California San Francisco, San Francisco, CA, USA. [4]Howard Hughes Medical Institute and Department of Molecular Biology, Massachusetts General Hospital, Harvard Medical School, Broad Institute, Cambridge, MA, USA. [5]Department of Cell Biology, University of Texas Southwestern Medical Center, Dallas, TX, USA. [6]Present address: Amgen Discovery Research, Cambridge, MA, USA. *e-mail: xiaochen.bai@utsouthwestern.edu; youxing.jiang@utsouthwestern.edu

© 2018 Springer Nature Limited. All rights reserved.

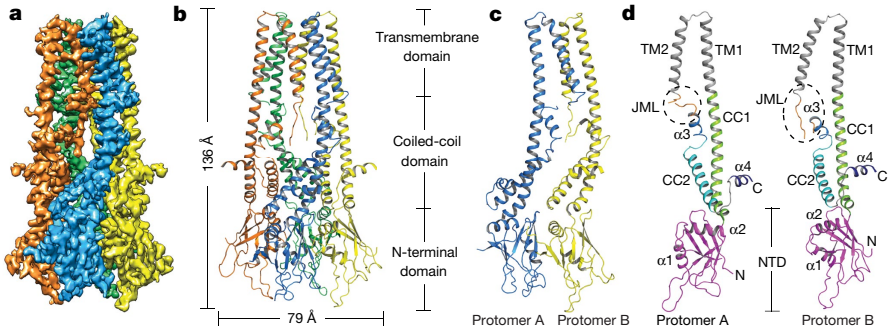


Fig. 1 | Cryo-EM structure of NfMCU. a, The overall cryo-EM map of NfMCU with density for saposin removed for clarity. **b**, Overall architecture of NfMCU, coloured by subunit and divided into major structural domains. **c**, Each NfMCU tetramer assembles as a dimer of dimers. Each dimer consists of two protomers adopting distinct conformations. **d**, Structural comparison of the two protomers in each MCU dimer. Dashed circles highlight the distinct conformation of the juxtamembrane loop (JML) in the two protomers.

oligomerization to maintain NfMCU in an artificial lipid bilayer that is proportional to the channel's smaller transmembrane domain. NfMCU reconstituted in saposin yielded a marked improvement in the structure to an overall resolution of 4.6 Å; however, several parts of the channel were poorly resolved, probably owing to inaccurate particle alignment stemming from the inherently low contrast of small particles. Finally, we used a Volta phase plate to image the NfMCU saposin complex, which ultimately allowed us to determine the structure to an overall resolution of 3.8 Å according to the gold-standard Fourier shell correlation (FSC) = 0.143 criterion (Extended Data Fig. 3, Extended Data Table 1 and Methods).

Consistent with our biochemical analysis of NfMCU in detergent, the EM map unambiguously showed tetrameric assembly of NfMCU (Fig. 1a). The map was of sufficient quality for model building of major parts of the protein (Extended Data Fig. 4 and Methods). However, parts of the N and C termini, as well as the long loop in the NTD, were disordered and were not modelled in the final structure (see Methods). The EM map also revealed clear electron density for the 81-residue saposin, which we modelled as poly-alanine (Extended Data Fig. 5). Six saposin molecules—each forming a C-shaped clamp oriented at an angle of approximately 45° to the central axis—encircle the lipid bilayer to stabilize NfMCU. Thus, our structure also reveals how saposin oligomerizes to maintain a membrane protein in lipid.

Overall architecture of NfMCU

NfMCU forms an elongated, three-tiered channel tetramer with each tier corresponding to a structural domain: the N-terminal domain (NTD), coiled-coil domain (CCD) and transmembrane domain (TMD) (Fig. 1b, Extended Data Fig. 6). Although NfMCU is tetrameric, it transitions from being four-fold symmetric at the TMD to two-fold symmetric in the NTD, which forms a dimer of dimers (Figs. 1c, 2a). Unlike the NMR structure of cMCU-ΔNTD, which lacks the NTD and adopts a constricted architecture, NfMCU is architecturally open, with large fenestrations in the space defined by the CCD (Fig. 1b, Extended Data Fig. 7a). Sitting atop the NTD, the CCD resembles pillars supporting the TMD. The tight packing among the four NTD subunits, as well as the exceptionally long CC1 helix that extends into TM1, dictates the spacing among the four CCDs, thereby, creating a highly solvent-accessible space below the ion conduction pathway.

There are several major differences between the structures of NfMCU and cMCU-ΔNTD (Extended Data Fig. 7a, b). In addition to their oligomeric states and overall architecture, the secondary structural elements between the two structures are also quite different. The CC1 and TM1 form a single extended helix in NfMCU but are comprised of five short, broken helices in cMCU-ΔNTD (Extended Data Fig. 7b). These structural differences are unlikely to be caused by the different origins of the two channels, as the CCD, TMD and selectivity filter are highly conserved among MCU orthologues[12] (Extended Data Fig. 6). We suspect that the missing NTD and the use of a denaturing detergent such as fos-choline to extract and purify protein from inclusion bodies in the NMR study contributed to these discrepancies.

Luminal soluble domains

The NTD and CCD constitute the luminal soluble domains of NfMCU. Despite the poor sequence conservation of the NTD among MCU orthologues[12] (Extended Data Fig. 6), the structures of the NTDs from NfMCU and human MCU are strikingly similar[14,15]. Consisting of seven β-strands (β1–7) and two helices (α1 and 2), each NTD forms a barrel enclosed by the seven-stranded β-sheet and the α1 helix, with the α2 helix sitting atop the barrel (Extended Data Fig. 7b). Four NTDs form a well-packed tetrameric ring in a dimer of dimers configuration with two-fold symmetry (Fig. 2a). Inter-subunit interactions are defined by two types of interface. Interface I involves primarily hydrogen bonding between main chain atoms on β5 of one subunit and β3 of a neighbouring subunit (Fig. 2b), whereas interface II is stabilized by hydrogen bonding between polar side chains (Fig. 2c). A similar dimerization at interface I is also observed in the crystal structure of $HsMCU_{NTD}$ (Extended Data Fig. 7c), thus highlighting the structural conservation between NfMCU and human MCU. The tetrameric assembly of the NTD implies that this domain has a role in maintaining the channel's architectural integrity. However, its functional importance is not entirely clear on the basis of studies of human MCU[14,16] and, thus, warrants further investigation.

The CCD consists of four helices—CC1, α3, CC2 and α4 (Fig. 1d, Extended Data Fig. 6). The first coiled-coil (CC1) forms an exceptionally long α-helix that curves with respect to the central axis as it extends from the NTD towards the TMD, transitioning seamlessly into TM1. A juxtamembrane loop (JML)—approximately 10 residues—and the α3

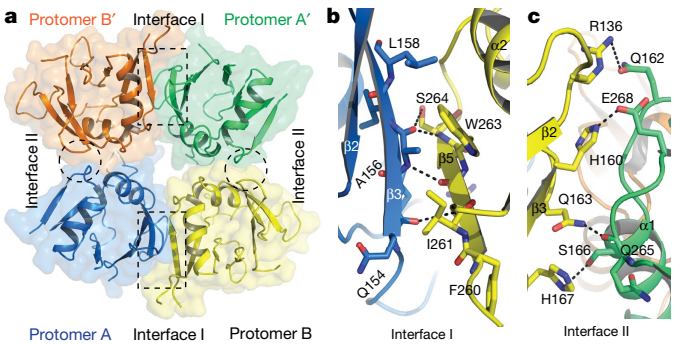


Fig. 2 | Channel assembly and subunit interfaces at the NTD. a, Top view of surface-rendered NfMCU showing inter-subunit interactions at the NTD. **b**, Detailed view of atomic interactions at interface I in NfMCU consisting of hydrogen bonds (dashed lines) mainly between main chain atoms. **c**, Atomic interactions at interface II consist of mainly side chain hydrogen bonds.

© 2018 Springer Nature Limited. All rights reserved.

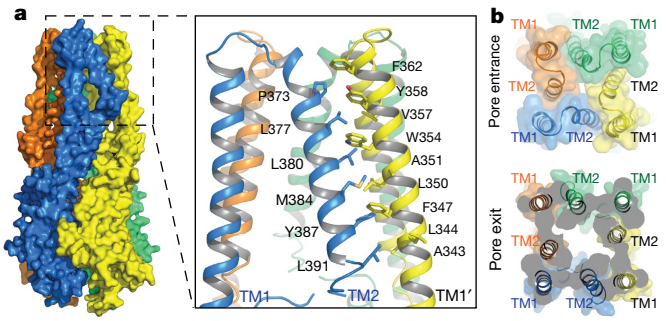


Fig. 3 | Channel assembly and subunit interfaces at the TMD. a, Left, surface-rendered model of NfMCU; right, zoomed-in view of the inter-subunit interactions at the TMD between TM2 of one subunit and TM1 of a neighbouring subunit. **b**, Top view of NfMCU at the pore entrance (top) and exit (bottom), highlighting the unique arrangement of the transmembrane helices in channel assembly and the widening of the pore moving down the ion conduction pathway.

helix connect the end of TM2 to the second coiled-coil (CC2) in two distinct conformations between two neighbouring subunits (Fig. 1d). In one subunit, the loop runs parallel to the membrane and the α3 helix runs almost parallel to CC1, while in the other subunit the loop extends nearly perpendicularly from the membrane and the α3 helix becomes more perpendicular to CC1. The CC2 helix runs nearly parallel to CC1 to form a coiled-coil and the CCD ends with the short α4 helix that runs perpendicularly to the coiled-coil, jutting away from the central axis (Fig. 1d).

Ion-conducting pore

The NfMCU pore is architecturally unique as compared to classical two-transmembrane tetrameric cation channels such as KcsA[21], whose outer TM1 helix wraps around the inner pore-forming TM2 helix, generating an inverted teepee-shaped ion conduction pore. Pillared by four TM1s, the transmembrane helices of NfMCU enclose a cone-shaped ion conduction pore with a wider opening at the luminal side (Figs. 3, 4a). Notably, the pore-forming TM2 makes very little contact with the TM1 from the same subunit, but rather interacts extensively with a neighbouring TM1 (Fig. 3a). Consequently, there is a wide gap between TM1 and TM2 within each subunit at the luminal half of the membrane, which we expect to be sealed by lipid.

NfMCU also lacks the canonical re-entrant pore loop and helix that define the selectivity filter of tetrameric cation channels; instead, the filter residues constitute the first helical turn of TM2. The conserved acidic residues of the DIME motif, which is $_{369}$DTME$_{372}$ in NfMCU, point towards the central axis and are separated by about one helical turn (Fig. 4). Four D369 residues encircle the cytosolic pore entrance of NfMCU with a diagonal distance of about 8.0 Å between carboxylate oxygen atoms. The side chains of E372 form the narrowest part of the filter, with the narrowest opposing carboxylate oxygens measuring about 4.5 Å (Fig. 4b). As NfMCU was maintained in 1 mM $CaCl_2$ throughout sample preparation, we also observed a strong electron density at the centre of the E372 ring, which we assigned as a Ca^{2+} chelated by two opposing E372s in a bidentate manner with ion-ligand distances between 2.2 and 2.5 Å, which are optimal for Ca^{2+} coordination (Fig. 4c, d). The protein packing around the filter region also provides insights into the evolutionary conservation of several residues close to the DIME motif. W368 intercalates between TM2 of each subunit with its indole ring oriented nearly vertically; its phenyl ring forms CH-π interactions with the highly conserved P373 in a stacked-like configuration[22,23], while the amino group of the pyrrole ring forms intra- and intermolecular hydrogen bonds with the carboxylate oxygen of E372 (Fig. 4c, d). Replacement of W368 or P373 with alanine resulted in non-functional channels (Fig. 4e). Thus, the region constituting the selectivity filter of NfMCU is tightly packed and conformationally rigid.

Beyond the selectivity filter, the pore opens into a large, mainly hydrophobic cavity that widens towards the base of the ion conduction pathway rather than constricting (as seen among canonical tetrameric cation channels) (Fig. 4a). Whether gating occurs within the ion-conducting pore is unclear and warrants further investigation. However, our in vitro $^{45}Ca^{2+}$ flux assay and bacterial Ca^{2+} uptake assay seem to suggest that NfMCU functions as a constitutively open channel, as the only condition required to stimulate Ca^{2+} uptake is an electrochemical gradient. Our speculation is also supported by the general consensus that Ca^{2+} activation of the human uniporter is mediated by MICU1 and MICU2 rather the channel itself[24–27].

Ca^{2+} permeation and inhibitor binding

The proximity of the two rings of acidic residues—D369s and E372s—within the selectivity filter of NfMCU suggests a two-site, single-file pore for selective Ca^{2+} permeation in MCU, which is conceptually similar to the classical ion conduction model for the voltage-gated Ca^{2+} channel (Ca_V)[28,29]. With a larger diameter, the ring of D369s at the cytosolic entrance encircles a low-affinity Ca^{2+} site (site 1) for hydrated ion. The site is likely to be non-specific, but to show a preference for Ca^{2+} given the highly negative charge of the ring. The narrow ring consisting of E372s forms the high-affinity, Ca^{2+}-selective site (site 2) with the acidic side chains directly chelating Ca^{2+}, analogous to the high-affinity Ca^{2+} site formed by the EEEE locus of Ca_V channels[30]. In addition, the structural constraint of E372 forces the carboxylate side chains to be near each other, making an empty site 2 energetically unfavourable owing to electrostatic repulsion. We suspect that, to mitigate this effect, site 2 is constitutively bound to Ca^{2+} with high affinity, which can be accomplished under physiological conditions because MCU has high affinity for Ca^{2+} ($\leq$2 nM)[3]. Thus, a potential mechanism for Ca^{2+}

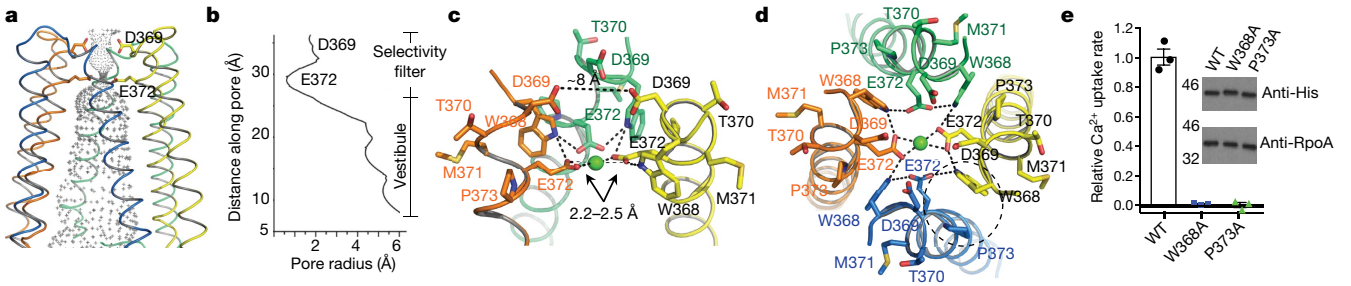


Fig. 4 | Ion conduction pore of NfMCU. a, Ribbon diagram of the NfMCU pore domain with the ion conduction pathway rendered as grey mesh and one pair of the acidic filter residues (D369 and E372) shown as sticks. **b**, Pore radius along the central axis. **c**, Side view of the selectivity filter region with the front subunit removed for clarity, highlighting the arrangement of the DIME motif and the approximate cross distance measurements between D369 residues and Ca^{2+} coordination distances for E372 (shown as dashed lines). Green sphere represents Ca^{2+}. **d**, Top view of the selectivity filter region. E372 forms a hydrogen bond network with W368. Ca^{2+} is coordinated by a pair of E372s in a bidentate manner. Dashed circle marks the inter-subunit packing between W368 and P373. **e**, Functional activity of NfMCU W368 and P373 mutants expressed in *Escherichia coli*. Data shown as mean ± s.e.m. (n = 3 independent experiments). Protein expression level was analysed by immunoblotting using *E. coli* RNA polymerase alpha (RpoA) as control (see Methods). The experiment was repeated three times independently with similar results.

© 2018 Springer Nature Limited. All rights reserved.

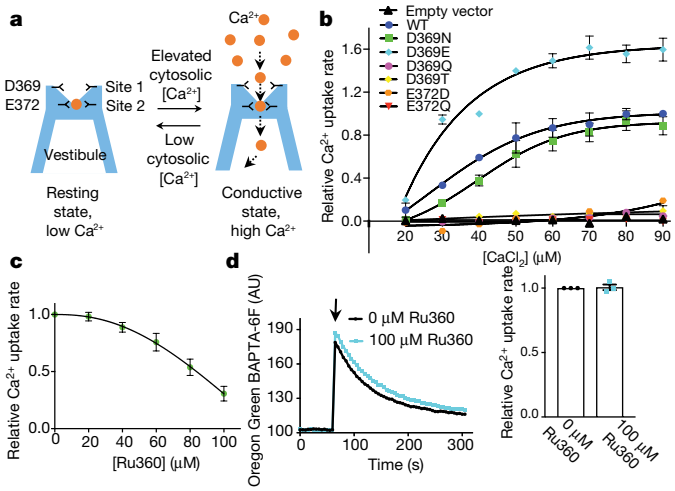


Fig. 5 | Mechanism of ion transport and Ru360 inhibition. a, Proposed model of calcium permeation in NfMCU. **b**, Concentration-dependent calcium uptake in *E. coli* expressing NfMCU filter mutations. Data are normalized to the calcium uptake rate of wild-type NfMCU at 90 μM $CaCl_2$. **c**, Normalized concentration-dependent Ru360 inhibition of *E. coli* expressing NfMCU D369N. **d**, Representative calcium uptake trace for *E. coli* expressing NfMCU D369E mutant with or without Ru360. Arrow indicates addition of 50 μM $CaCl_2$ to the reaction. Bar graph shows the rate of calcium uptake normalized to the uptake rate without Ru360 in the reaction. Data in **b**, **c** shown as mean ± s.e.m. ($n = 3$ independent experiments).

transport in MCU is likely to follow a similar model to that proposed for Ca_V channels (Fig. 5a): Ca^{2+} binds constitutively to site 2 under physiological conditions and blocks the permeation of Na^+ or K^+. With increasing cytosolic Ca^{2+} concentration, Ca^{2+} gradually occupies site 1 and, thereby, provide the necessary electrostatic repulsion between sites 1 and 2 to lower the affinity of the site 2 Ca^{2+}, allowing it to be readily knocked off and to move into the water-filled vestibule directly below the filter. This Ca^{2+} permeation cycle repeats as the Ca^{2+} from site 1 hops to site 2, freeing site 1 to bind a new Ca^{2+}. The presence of the low-affinity site external to the high-affinity site may also explain the inward rectification of Ca^{2+} permeation observed in MCU.

The two-site model implies that the high affinity site 2 is essential for Ca^{2+} selectivity while the low affinity site 1 is less specific but necessary for efficient Ca^{2+} permeation. We tested the roles of the acidic residues at both sites for Ca^{2+} permeation in NfMCU using our bacterial uptake assay. We first examined the effect of neutralizing the two acidic residues individually (D369N and E372Q). Consistent with observations in human MCU[4,16], the E372Q mutant no longer conducts Ca^{2+}, and the D369N mutant exhibited a slight reduction in Ca^{2+} uptake (Fig. 5b, Extended Data Fig. 8). We also investigated whether disruption of the ion coordination distance had any effect on channel function. Notably, a D369E mutant showed about 50% greater Ca^{2+} uptake compared to the wild type, whereas the E372D mutant remained nonfunctional. A similar D261E mutation in human MCU—equivalent to the D369E mutation in NfMCU—showed a slight reduction in calcium uptake[16]. Mutation of D369 to threonine or glutamine yielded nonfunctional channels. These data confirm that site 2 (E372), which is essential for specific Ca^{2+} binding, is highly sensitive to mutations, whereas the non-specific site 1 (D369) is more tolerant to mutations.

The structure of NfMCU also offers an opportunity to probe the mechanism of binding of Ru360, an analogue of ruthenium red that specifically blocks Ca^{2+} uptake by MCU. In human MCU, a serine (S259) two residues before the DIME motif was initially shown to be important for Ru360 binding[4], as an alanine mutation at this site conferred resistance to the drug while still allowing calcium transport. In NfMCU, the equivalent residue is a glycine. Studies on cMCU-ΔNTD identified the aspartate of the DIME motif as a second key residue for Ru360 binding[31], which we suspect is the primary Ru360 binding site for NfMCU because the DIME motif is conserved across all MCU orthologues. Neutralizing the aspartate (D369N) weakened Ru360 blocking, whereas maintaining the charge but reducing the pore diameter at site 1 (D369E) eliminated Ru360 blocking (Fig. 5c, d). This result confirms that D369 in NfMCU forms the primary blocking site for Ru360, whose binding is dictated by the charge as well as the pore size at site 1.

Conclusion

Thus, the structure of NfMCU provides a model of an intact mitochondrial calcium uniporter determined to near-atomic resolution, providing new insights into channel assembly and function. Contrary to the structure of cMCU-ΔNTD, NfMCU functions as a tetrameric calcium channel, whose pore is poised for rapid ion conduction. In essence, the channel provides a rather finite barrier measuring about 6 Å—the distance separating the filter aspartate and glutamate—that Ca^{2+} must traverse from the intermembrane space to the mitochondrial matrix. The combination of two ion binding sites and a huge electrochemical gradient across the inner mitochondrial membrane provide the necessary driving force for MCU to selectively conduct Ca^{2+}. It is notable that early studies of calcium transport demonstrated that mitochondria from diverse eukaryotes have uniporter activity[11]. However, the kinetics and apparent binding affinities of calcium uptake by fungal mitochondria are appreciably different from those of vertebrate mitochondria[11]. While NfMCU reveals the architectural arrangement of the basic channel component of the uniporter, further study is needed to address fundamental questions about metazoan uniporter function, such as how the metazoan MCU evolved to require EMRE for ion conduction and how MICU1 and MICU2 confer Ca^{2+}-dependent gating to the uniplex.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0333-6.

Received: 5 January 2018; Accepted: 15 May 2018;
Published online 11 July 2018.

1. Kamer, K. J. & Mootha, V. K. The molecular era of the mitochondrial calcium uniporter. *Nat. Rev. Mol. Cell Biol.* **16**, 545–553 (2015).
2. Berridge, M. J., Bootman, M. D. & Roderick, H. L. Calcium signalling: dynamics, homeostasis and remodelling. *Nat. Rev. Mol. Cell Biol.* **4**, 517–529 (2003).
3. Kirichok, Y., Krapivinsky, G. & Clapham, D. E. The mitochondrial calcium uniporter is a highly selective ion channel. *Nature* **427**, 360–364 (2004).
4. Baughman, J. M. et al. Integrative genomics identifies MCU as an essential component of the mitochondrial calcium uniporter. *Nature* **476**, 341–345 (2011).
5. De Stefani, D., Raffaello, A., Teardo, E., Szabò, I. & Rizzuto, R. A forty-kilodalton protein of the inner membrane is the mitochondrial calcium uniporter. *Nature* **476**, 336–340 (2011).
6. Sancak, Y. et al. EMRE is an essential component of the mitochondrial calcium uniporter complex. *Science* **342**, 1379–1382 (2013).
7. Plovanich, M. et al. MICU2, a paralog of MICU1, resides within the mitochondrial uniporter complex to regulate calcium handling. *PLoS One* **8**, e55785 (2013).
8. Perocchi, F. et al. MICU1 encodes a mitochondrial EF hand protein required for Ca(2+) uptake. *Nature* **467**, 291–296 (2010).
9. Tsai, M.-F. et al. Dual functions of a small regulatory subunit in the mitochondrial calcium uniporter complex. *eLife* **5**, e15545 (2016).
10. Kovács-Bogdán, E. et al. Reconstitution of the mitochondrial calcium uniporter in yeast. *Proc. Natl Acad. Sci. USA* **111**, 8985–8990 (2014).
11. Carafoli, E. & Lehninger, A. L. A survey of the interaction of calcium ions with mitochondria from different tissues and species. *Biochem. J.* **122**, 681–690 (1971).
12. Bick, A. G., Calvo, S. E. & Mootha, V. K. Evolutionary diversity of the mitochondrial calcium uniporter. *Science* **336**, 886 (2012).
13. Song, J., Liu, X., Zhai, P., Huang, J. & Lu, L. A putative mitochondrial calcium uniporter in A. fumigatus contributes to mitochondrial Ca^{2+} homeostasis and stress responses. *Fungal Genet. Biol.* **94**, 15–22 (2016).
14. Lee, Y. et al. Structure and function of the N-terminal domain of the human mitochondrial calcium uniporter. *EMBO Rep.* **16**, 1318–1333 (2015).
15. Kishnani, P. S. et al. Immune response to enzyme replacement therapies in lysosomal storage diseases and the role of immune tolerance induction. *Mol. Genet. Metab.* **117**, 66–83 (2016).
16. Oxenoid, K. et al. Architecture of the mitochondrial calcium uniporter. *Nature* **533**, 269–273 (2016).

© 2018 Springer Nature Limited. All rights reserved.

17. Wang, L. et al. Structural and mechanistic insights into MICU1 regulation of mitochondrial calcium uptake. *EMBO J.* **33**, 594–604 (2014).
18. Logan, C. V. et al. Loss-of-function mutations in MICU1 cause a brain and muscle disorder linked to primary alterations in mitochondrial calcium signaling. *Nat. Genet.* **46**, 188–193 (2014).
19. Gunter, T. E. & Pfeiffer, D. R. Mechanisms by which mitochondria transport calcium. *Am. J. Physiol.* **258**, C755–C786 (1990).
20. Frauenfeld, J. et al. A saposin-lipoprotein nanoparticle system for membrane proteins. *Nat. Methods* **13**, 345–351 (2016).
21. Doyle, D. A. et al. The structure of the potassium channel: molecular basis of K+ conduction and selectivity. *Science* **280**, 69–77 (1998).
22. Biedermannova, L., E Riley, K., Berka, K., Hobza, P. & Vondrasek, J. Another role of proline: stabilization interactions in proteins and protein complexes concerning proline and tryptophane. *Phys. Chem. Chem. Phys.* **10**, 6350–6359 (2008).
23. Zondlo, N. J. Aromatic-proline interactions: electronically tunable CH/π interactions. *Acc. Chem. Res.* **46**, 1039–1049 (2013).
24. Mallilankaraman, K. et al. MICU1 is an essential gatekeeper for MCU-mediated mitochondrial Ca^{2+} uptake that regulates cell survival. *Cell* **151**, 630–644 (2012).
25. Kamer, K. J. & Mootha, V. K. MICU1 and MICU2 play nonredundant roles in the regulation of the mitochondrial calcium uniporter. *EMBO Rep.* **15**, 299–307 (2014).
26. Patron, M. et al. MICU1 and MICU2 finely tune the mitochondrial Ca^{2+} uniporter by exerting opposite effects on MCU activity. *Mol. Cell* **53**, 726–737 (2014).
27. Liu, J. C. et al. MICU1 serves as a molecular gatekeeper to prevent *in vivo* mitochondrial calcium overload. *Cell Reports* **16**, 1561–1573 (2016).
28. Almers, W. & McCleskey, E. W. Non-selective conductance in calcium channels of frog muscle: calcium selectivity in a single-file pore. *J. Physiol. (Lond.)* **353**, 585–608 (1984).
29. Hess, P. & Tsien, R. W. Mechanism of ion permeation through calcium channels. *Nature* **309**, 453–456 (1984).
30. Heinemann, S. H., Terlau, H., Stühmer, W., Imoto, K. & Numa, S. Calcium channel characteristics conferred on the sodium channel by single mutations. *Nature* **356**, 441–443 (1992).
31. Cao, C., Wang, S., Cui, T., Su, X.-C. & Chou, J. J. Ion and inhibitor binding of the double-ring ion selectivity filter of the mitochondrial calcium uniporter. *Proc. Natl Acad. Sci. USA* **114**, E2846–E2851 (2017).

Acknowledgements We thank J. Guo and Q. Chen for technical assistance and manuscript preparation, and J. Frauenfeld (Salipro Biotech) for the recombinant saposin construct. Negative-stained EM data acquisition and sample optimization for cryo-EM were performed at the University of California San Francisco; we thank M. Braunfeld for supervision of the microscopes. Phase plate cryo-EM data were collected at the University of Texas Southwestern Medical Center Cryo-EM Facility, which is funded by the CPRIT Core Facility Support Award RP170644; we thank D. Nicastro and Z. Chen for technical support and facility access. This work was supported in part by the Howard Hughes Medical Institute (to Y.J., V.K.M. and Y.C.), the National Institutes of Health (grant GM079179 to Y.J.; grants R01GM098672, S10OD020054, and S10OD021741 to Y.C.), the Welch Foundation (Grant I-1578 to Y.J.), the Cancer Prevention and Research Initiative of Texas (to X.B.) and the Virginia Murchison Linthicum Scholar in Medical Research fund (to X.B.).

Reviewer information *Nature* thanks M. Prakriya, D. Stokes and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions N.X.N., V.K.M. and Y.J. conceived the study. N.X.N., J.-P.A., Y.J., Y.C. and X.B. designed the experiments and analysed the data. N.X.N., J.-P.A., C.L., Y.Y., W.Z., and X.B. generated key materials and executed the experiments. N.X.N. and Y.J. wrote the manuscript with input from J.-P.A., Y.C., V.K.M. and X.B. Y.J. supervised the project and revised the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0333-6.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0333-6.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to X.B. or Y.J.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Expression and purification of NfMCU. The construct for NfMCU (NCBI accession number: XP_001266985.1) was codon optimized for expression in *E. coli* and cloned into the pCOLD II vector[32]. For functional assay, the full-length protein with only the mitochondrial targeting sequence removed (Extended Data Fig. 5) was modified to contain an N-terminal 6x-His tag and a thrombin cleavage site (LVPRGS) and then reinserted back into the pCOLD II vector (Takara) using the restriction sites NdeI/XhoI. For structural studies, the NfMCU construct used in the function assay was further truncated by an additional 10 residues each from the N and C termini and then cloned into the pCOLD II vector using the same restriction sites. Although no significant functional difference was observed between the full-length protein and the truncated channel (Extended Data Fig. 1b), the truncated construct yields a significantly more monodispersed profile on gel filtration and was subsequently used for structural studies (Extended Data Fig. 1a). All single-site mutants were generated using the Quikchange Site-Directed Mutagenesis Kit (Agilent) and confirmed by DNA sequencing.

E. coli (BL21 DE3) were freshly transformed, plated onto LB/ampicillin (LB/Amp) agar, and grown overnight. Colonies were harvested directly from LB/Amp plates to seed LB media (100 μg/mL ampicillin) and grown in shaker flasks at 37 °C in an orbital shaker until the cells reached an $OD_{600} = 0.37–0.4$, at which point they were transferred to another orbital shaker pre-chilled to 15 °C and allowed to grow for another hour before protein expression was induced by addition of IPTG to a final concentration of 0.4 mM. Cells were then grown at 15 °C for 16–18 h before being harvested. The cell pellets were flash frozen in liquid nitrogen and stored at −80 °C.

To purify NfMCU, the cell pellets were resuspended in lysis buffer (50 mM Tris pH 8, 300 mM NaCl, 1 mM $CaCl_2$), supplemented with protease inhibitors (1 μg/mL each of DNase I, pepstatin, leupeptin and aprotinin, and 1 mM PMSF), and then homogenized on ice. To remove inclusion bodies and misfolded protein, the homogenate was centrifuged at 10,000 × g for 20 min at 4 °C. The supernatant was harvested and NfMCU was extracted with 1% (w:v) n-dodecyl-β-D-maltopyranoside (DDM, Anatrace) by gentle agitation for 1 h on ice. After extraction, the supernatant was collected following a 30-min centrifugation at 21,000 × g at 4 °C and then incubated with Ni-NTA resin (Qiagen) with gentle agitation on ice for 1 h. The resin was collected on a disposable gravity column (Bio-Rad), washed with wash buffer A (lysis buffer + 1 mM DDM) for 3 column volumes and then with wash buffer B (lysis buffer + 1 mM DDM + 30 mM imidazole) for 10 column volumes. NfMCU was eluted with wash buffer C (lysis buffer + 1 mM DDM + 300 mM imidazole). Thrombin was added at a ratio of 1 unit of thrombin per 1 L culture of bacterial cell pellet; the sample was allowed to incubate at 20 °C, overnight, to remove the His-tag. NfMCU was then concentrated and further purified by size exclusion chromatography on a Superdex 200 10/300 GL column (GE Heathcare) pre-equilibrated with buffer D (20 mM HEPES pH 7.5, 300 mM NaCl, 1 mM $CaCl_2$, 1 mM DDM). The main peak (eluting around 11 mL) was then collected for further biochemical and structural analysis.

Reconstitution of NfMCU into saposin. We were unsuccessful in reconstituting NfMCU following the original protocol[20], likely due to the recommended step of heat shocking the sample during reconstitution. We thus made some modifications to the protocol, which we believe is better suited for more sensitive or less stable membrane proteins. Our reaction consists of the following components mixed at a molar ratio of 1:3:10 (NfMCU: saposin: *E. coli* total lipids). As a substitution for the recommended heat-shock step[20], saposin was destabilized by incubation with sodium cholate (15 mM final concentration) with rotation at 4 °C for 1 h. Purified NfMCU (concentrated to 100 μM) was then added to the reaction and allowed to rotate at 4 °C for 1 h before addition of BioBeads (BioRad) to remove detergents. The reaction was incubated with BioBeads for 3–4 h at 4 °C, and BioBeads exchange was performed a total of 3 times with each incubation step lasting 3–4 h at 4 °C. The reconstituted NfMCU saposin complex was separated from empty saposin complex and aggregates by gel filtration chromatography on a Superdex 200 10/300 GL column (GE Heathcare) pre-equilibrated with buffer E (20 mM HEPES pH 7.5, 300 mM NaCl, 1 mM $CaCl_2$, 2% glycerol). The main peak (eluting around 11 mL) was then collected for structural analysis.

EM data acquisition. Prior to freezing grids, the NfMCU saposin complex (concentrated to 0.6 mg/mL) was incubated with 50 μM Ru360 on ice for 30 min. The cryo-EM grids were prepared by applying 3 μL of the NfMCU-saposin-Ru360 complex to a glow-discharged Quantifoil R1.2/1.3 300-mesh gold holey carbon grid (Quantifoil, Micro Tools GmbH, Germany) and blotted for 4.0 s under 100% humidity at 4 °C before being plunged into liquid ethane using a Mark IV Vitrobot (FEI). Micrographs were acquired on a Titan Krios microscope (FEI) operated at 300 kV with a K2 Summit direct electron detector (Gatan), using a slit width of 20 eV on a GIF-Quantum energy filter as well as a Volta phase plate following the same approach for phase plate data collection as previously described[33]. EPU software (FEI) was used for automated data collection following standard FEI procedure. A calibrated magnification of 59,523× was used for imaging, yielding a pixel size of 0.84 Å on images. The defocus was set at −0.5 μm. Each micrograph was dose-fractionated to 30 frames under a dose rate of 4 e^-/pixel/s, with a total exposure time of 12 s, resulting in a total dose of about 60 $e^-/Å^2$.

Image processing. For the NfMCU phase plate data set, motion correction was performed using the MotionCorr2 program[34], and the CTF parameters of the micrographs were estimated using the GCTF program[35]. All other steps of image processing were performed using RELION[36]. Initially, about 1,000 particles were manually picked from a few micrographs. Class averages representing projections of NfMCU in different orientations were selected from the 2D classification of the manually picked particles, and used as templates for automated particle picking from the full data set of 2,470 micrographs. The 526,554 extracted particles were binned 3 times and subjected to two rounds of 2D classification, and a total of 355,929 particles were finally selected for 3D classification using the initial model generated by RELION as the reference. Two of the 3D classes showed good secondary structural features and their particles were selected, combined and re-extracted into the original pixel size of 0.84 Å. After 3D refinement with *C2* symmetry imposed and particle polishing, the resulting 3D reconstructions from 128,231 particles yielded an EM-map with a resolution of 3.8 Å. The transmembrane domain showed poor density, indicating a local structural heterogeneity. Therefore, we performed a focused 3D classification with density subtraction in order to improve the density of the transmembrane domain[37]. In this approach, the density corresponding to the soluble domains of the channel as well as the belt-like density from saposin was subtracted from the original particles. The subsequent 3D classification on the modified particles was carried out by applying a mask around the transmembrane domain and having all the orientations fixed at the value determined in the initial 3D refinement. After this round of classification, one class (83,343 particles) with better density in the transmembrane domain was selected for 3D refinement, yielding an EM map of 3.8 Å in which the entire transmembrane domain can be modelled. All resolutions were estimated by applying a soft mask around the protein density and the gold-standard FSC = 0.143 criterion. ResMap[38] was used to calculate the local resolution map.

Model building, refinement, and validation. De novo atomic model building was conducted in Coot[39]. Since the crystal structure of the NTD of human MCU[14] (PDB: 4XTB) fits nicely in the EM map, it was used as a guide to build the NTD of NfMCU. Amino acid assignment was achieved based on the clearly defined density for bulky residues (Phe, Trp, Tyr, and Arg). Models were refined against summed maps using phenix.real_space_refine[40], with secondary structure restraints applied. The model was validated using previously described methods to avoid overfitting[41,42]. The 3.8 Å EM density map of NfMCU allowed us to construct a model containing the following residues: chain A, B: 124–200, 258–399, and 406–461; chain C, D: 124–196, 259–396, and 404–461. Although we added Ru360 in our sample with the hope of seeing how it blocks the channel pore, its density was not visible in the structure. One main reason could be the symmetry mismatch. As the MCU pore entrance has fourfold symmetry, Ru360 could bind to the pore in four different orientations, causing its density to be averaged out. The statistics for the model's geometries were generated using MolProbity[43]. Pore radii were calculated using the HOLE program[44]. All the figures were prepared in PyMol[45] or Chimera[46].

Bacterial calcium uptake assay. *E. coli* expressing NfMCU were grown as described above with the exception that protein induction was allowed to proceed for 16 h, at which time the cells were harvested, centrifuged at 4000 × g for 10 min at 4 °C, and the media was discarded. To remove residual media, the cell pellet was resuspended in ice-cold buffer (50 mM HEPES pH 7.4, 150 mM KCl) and then centrifuged again (4000 × g for 10 min at 4 °C) and the supernatant was discarded. To prepare the bacteria for calcium uptake measurements, the bacteria were treated as though they were mitochondria and prepared as previously described for the purification of mitochondria from yeast[47] but with slight modifications. To remove residual Ca^{2+}, the cells were resuspended to a final concentration of 10 mg/mL in ice-cold bacteria prep buffer (0.6 M D-mannitol, 10 mM KH_2PO_4, 20 mM HEPES pH 7.4, 5 mg/mL bovine serum albumin) plus 1 mM EDTA and allowed to shake gently on ice for 30 min. To replenish any Mg^{2+} that was chelated by EDTA treatment, 2 mM $MgCl_2$ was then added into the cell resuspension followed by gentle shaking on ice for another 30 min. The cells were then centrifuged (4000 × g for 10 min at 4 °C) and the supernatant was discarded. The cells were centrifuged again (4000 × g for 3 min at 4 °C) and residual buffer was removed with a pipette. The wet weight of the bacteria was measured with a digital balance (typical yield is ~80 mg of bacteria per 30 mL culture of bacteria). The cells were then resuspended in ice-cold reaction buffer (0.6 M D-mannitol, 20 mM HEPES pH 6.8, 10 μM EGTA, 0.5 mg/mL bovine serum albumin) to a final concentration of 10 mg/mL and were used immediately for calcium uptake measurement.

To measure calcium uptake, we relied on the highly negative membrane potential across the plasma membrane of *E. coli*[48] and used the fluorescent calcium indicator Oregon Green BAPTA-6F (Life Technologies), which was added to a final concentration of 1 μM (from aliquots of a 500 μM stock kept at −20 °C) to cells resuspended in reaction buffer, to monitor the Ca^{2+} concentration changes

© 2018 Springer Nature Limited. All rights reserved.

over time using fluorescence signal—a conventional approach for monitoring mitochondrial calcium uptake in the field[4,5]. For calcium uptake measurements, 100 μL of cell suspension was added to a 96-well plate (Corning) and fluorescence was monitored in a Molecular Devices SpectraMax M3 plate reader (excitation/emission: 485 nm/535 nm). Baseline fluorescence was measured for all samples at the start of each reaction for 60 s followed by addition of 5 μL $CaCl_2$ stock to the 100 μL reaction in any experiment to reach the desired concentration. With the exception of the experiment in Fig. 1A, where the bacteria were pulsed 3 times with 20 μM $CaCl_2$, all measurements were from samples pulsed only once with the concentration of $CaCl_2$ indicated in the figure legend. Fluorescence was monitored for 4 min to assess calcium uptake.

For the inhibition experiments (either with $GdCl_3$ or Ru360), the inhibitor was added to 1000 μL of cell suspension to the highest desired concentration of inhibitor for the experiment and this cell suspension was then used as a stock to dilute into a cell suspension without any inhibitor added (to a final volume of 100 μL) to obtain a series of cells in the desired inhibitor concentration. The samples were then pulsed with 50 μM $CaCl_2$ and fluorescence was monitored for 4 min.

To ensure protein expression was comparable among the NfMCU mutants tested in our uptake assay, western blot analysis was performed on each sample used for each experiment. Extended Data Fig. 8 shows representative western blots of an experimental trial using the NfMCU mutants referred to in this study. Mouse anti-His antibodies (Qiagen) was used to detect NfMCU expression and mouse anti-*E. coli* RNA polymerase alpha or anti-RpoA (Santa Cruz Biotechnology) was used as a loading control. HRP-conjugated sheep anti-mouse antibodies were used as the secondary antibodies (GE Healthcare).

All data were analysed in GraphPad Prism 7 (GraphPad Software, Inc.). To obtain the rate of calcium uptake, the linear phase of the uptake measurement—the first 45–60 s of the reaction following $CaCl_2$ administration—was determined by fitting the data to a linear regression equation in GraphPad Prism; the slope of the fit is taken as the rate of calcium uptake. Further analysis such as normalization are indicated in the figure legend. To obtain the IC_{50} for Ru360, the relative calcium uptake rates were fitted to a dose–response curve in GraphPad Prism.

Radioactive ^{45}Ca flux assay. In light of the contradiction of the oligomeric state between NfMCU and cMCU-ΔNTD, we designed an in vitro functional assay by reconstituting our purified tetrameric channel into liposomes. Unable to obtain currents of NfMCU by electrophysiology—probably due to the channel's small conductance—we performed a radioactive $^{45}Ca^{2+}$ flux assay on proteoliposomes loaded with 150 mM KCl.

NfMCU was reconstituted into lipid vesicles using a similar protocol as described[49]. *E. coli* total lipid (Avanti) was prepared following standard protocol and then solubilized in dialysis buffer (50 mM HEPES pH 7.4, 150 mM KCl) supplemented with 40 mM n-decyl-β-D-maltopyranoside (DM, Anatrace). Purified NfMCU in DDM (at ~1 mg/mL) was mixed with the DM-solubilized *E. coli* total lipid at a ratio of 1:100 (w:w, protein:lipids). NfMCU and lipid mixture were incubated with gentle rotation at 4 °C for 1 h and then dialysed (using dialysis tubing with a molecular weight cutoff of 12,000–14,000 Da (Spectra/Por) against 1 L of dialysis buffer at 4 °C to remove detergent. Fresh dialysis buffer (1 L each time) was exchanged every 18–20 h for a total of 3 exchanges. Following dialysis, NfMCU proteoliposomes were divided into 100 μL aliquots, flash frozen in liquid nitrogen, and stored at −80 °C. For all flux assays, three separate batches of bacteria were grown to express NfMCU. The protein was then purified individually and used immediately for reconstitution into liposomes.

The $^{45}Ca^{2+}$ flux assay was performed following the same procedures as previously described but with slight modifications[49]. NfMCU proteoliposomes were thawed and sonicated in a bath sonicator for 60 s at 15 s intervals before the assay. 100 μL samples were passed through a pre-spun Sephadex G-50 fine gel filtration column (1.5 mL bed volume in a 5 mL disposable spin column) swollen in flux buffer (50 mM HEPES pH 7.4, 150 mM KCl or NMDG). The resulting 160 μL of proteoliposome eluent was collected and added to 300 μL of reaction buffer (flux buffer + 0.2 μM valinomycin + 83 μM $^{45}CaCl_2$) to initiate the reaction. At various time points, 75 μL of this reaction mixture was passed through a pre-spun Sephadex G-50 fine gel filtration column as previously described to remove extra-liposomal $^{45}Ca^{2+}$ and stop the flux. The eluent was mixed with 10 mL scintillation cocktail and its radioactivity was measured in a scintillation counter.

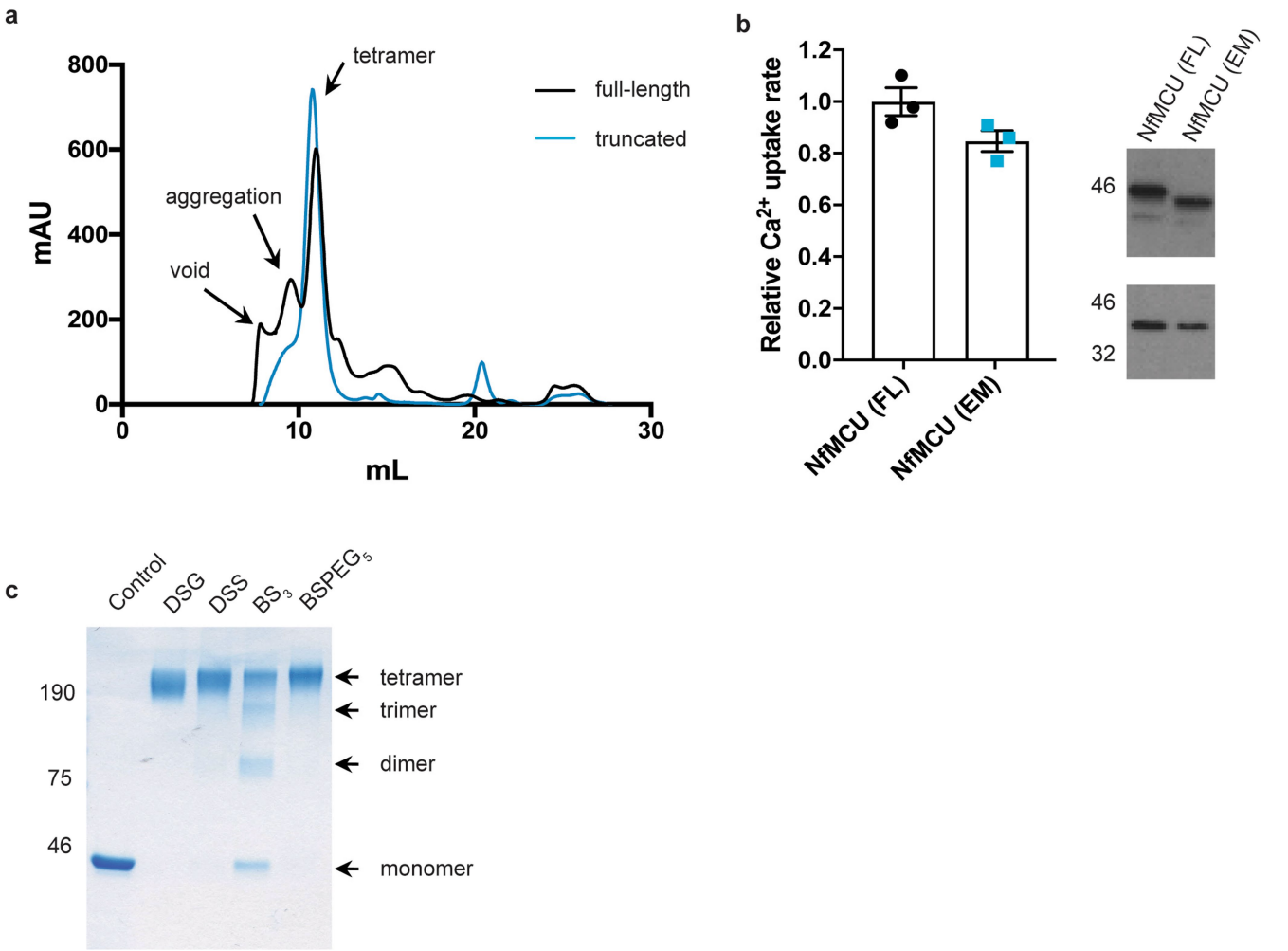
For the Ru360 inhibition experiments, NfMCU proteoliposomes in flux buffer (50 mM HEPES pH 7.4, 150 mM NMDG) were mixed with reaction buffer containing varying concentrations of Ru360 (dissolved in NMDG flux buffer + 0.2 μM valinomycin + 83 μM $^{45}CaCl_2$). The reaction was allowed to proceed for 30 min, after which the reaction was stopped by passing 75 μL of the reaction through a pre-spun Sephadex G-50 fine gel filtration column as previously described. To determine the relative calcium uptake by NfMCU proteoliposomes, the scintillation count for each sample was normalized to the sample without any Ru360 in the reaction.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. The cryo-EM density maps of the NfMCU have been deposited in the Electron Microscopy Data Bank under accession numbers EMD-7826 (without saposin) and EMD-7828 (with saposin, low-pass filtered to 5.0 Å resolution). Atomic coordinates have been deposited in the Protein Data Bank under accession numbers 6D7W (without saposin) and 6D80 (with saposin modelled as poly-A peptide). Source data for Figs. 4e, 5b–d, Extended Data Figs. 1b, 2 are available in the online version of the paper. The gel source data for Fig. 4e, Extended Data Figs. 1b, c, 8 are available in Supplementary Fig. 1.

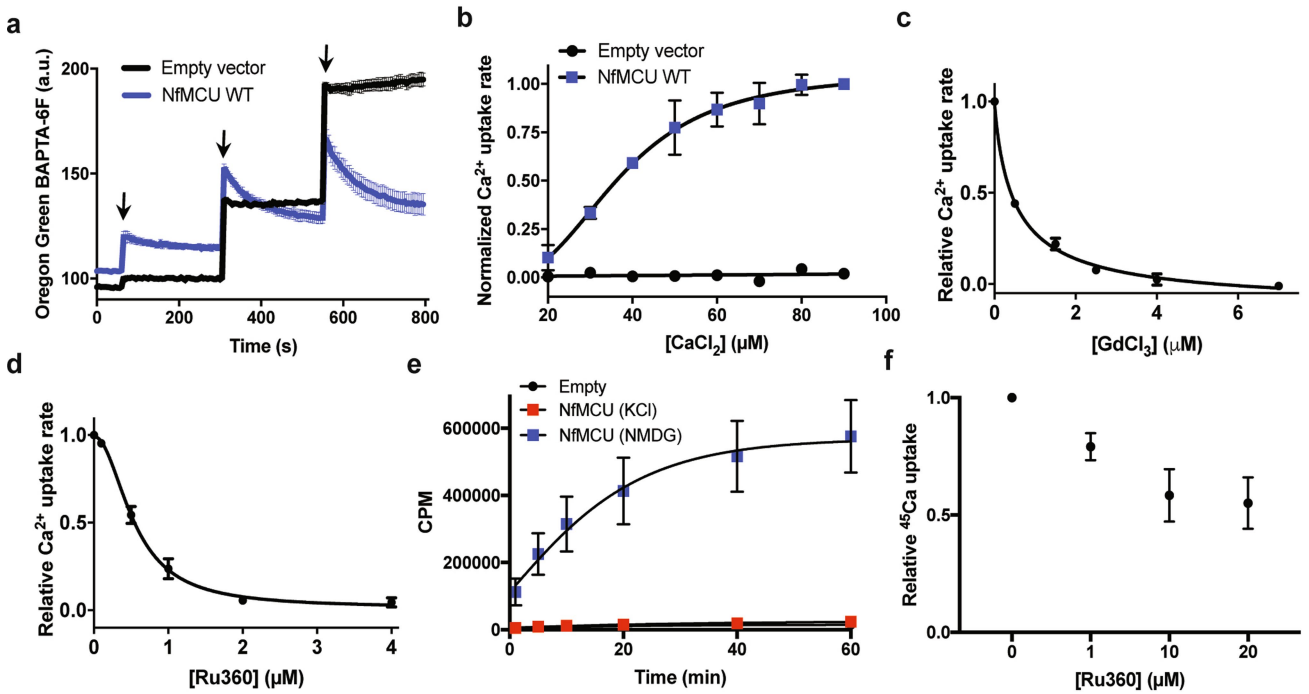
32. Pei, J., Kim, B. H. & Grishin, N. V. PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* **36**, 2295–2300 (2008).
33. Qing, G. et al. Cold-shock induced high-yield protein production in *Escherichia coli*. *Nat. Biotechnol.* **22**, 877–882 (2004).
34. Danev, R., Tegunov, D. & Baumeister, W. Using the Volta phase plate with defocus for cryo-EM single particle analysis. *eLife* **6**, e23006 (2017).
35. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
36. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
37. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
38. Bai, X. C., Rajendra, E., Yang, G., Shi, Y. & Scheres, S. H. Sampling the conformational space of the catalytic subunit of human γ-secretase. *eLife* **4**, e11182 (2015).
39. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).
40. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
41. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
42. Amunts, A. et al. Structure of the yeast mitochondrial large ribosomal subunit. *Science* **343**, 1485–1489 (2014).
43. DiMaio, F., Zhang, J., Chiu, W. & Baker, D. Cryo-EM model validation using independent map reconstructions. *Protein Sci.* **22**, 865–868 (2013).
44. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 12–21 (2010).
45. Smart, O. S., Neduvelil, J. G., Wang, X., Wallace, B. A. & Sansom, M. S. HOLE: a program for the analysis of the pore dimensions of ion channel structural models. *J. Mol. Graph.* **14**, 354–360, 376 (1996).
46. Schrödinger, L. The PyMOL Molecular Graphics System, Version 1.8. (2015).
47. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
48. Yaffe, M. P. Analysis of mitochondrial function and assembly. *Methods Enzymol.* **194**, 627–643 (1991).
49. Lo, C. J., Leake, M. C., Pilizota, T. & Berry, R. M. Nonequivalence of membrane voltage and ion-gradient as driving forces for the bacterial flagellar motor at low load. *Biophys. J.* **93**, 294–302 (2007).
50. Heginbotham, L., Kolmakova-Partensky, L. & Miller, C. Functional reconstitution of a prokaryotic K+ channel. *J. Gen. Physiol.* **111**, 741–749 (1998).

© 2018 Springer Nature Limited. All rights reserved.



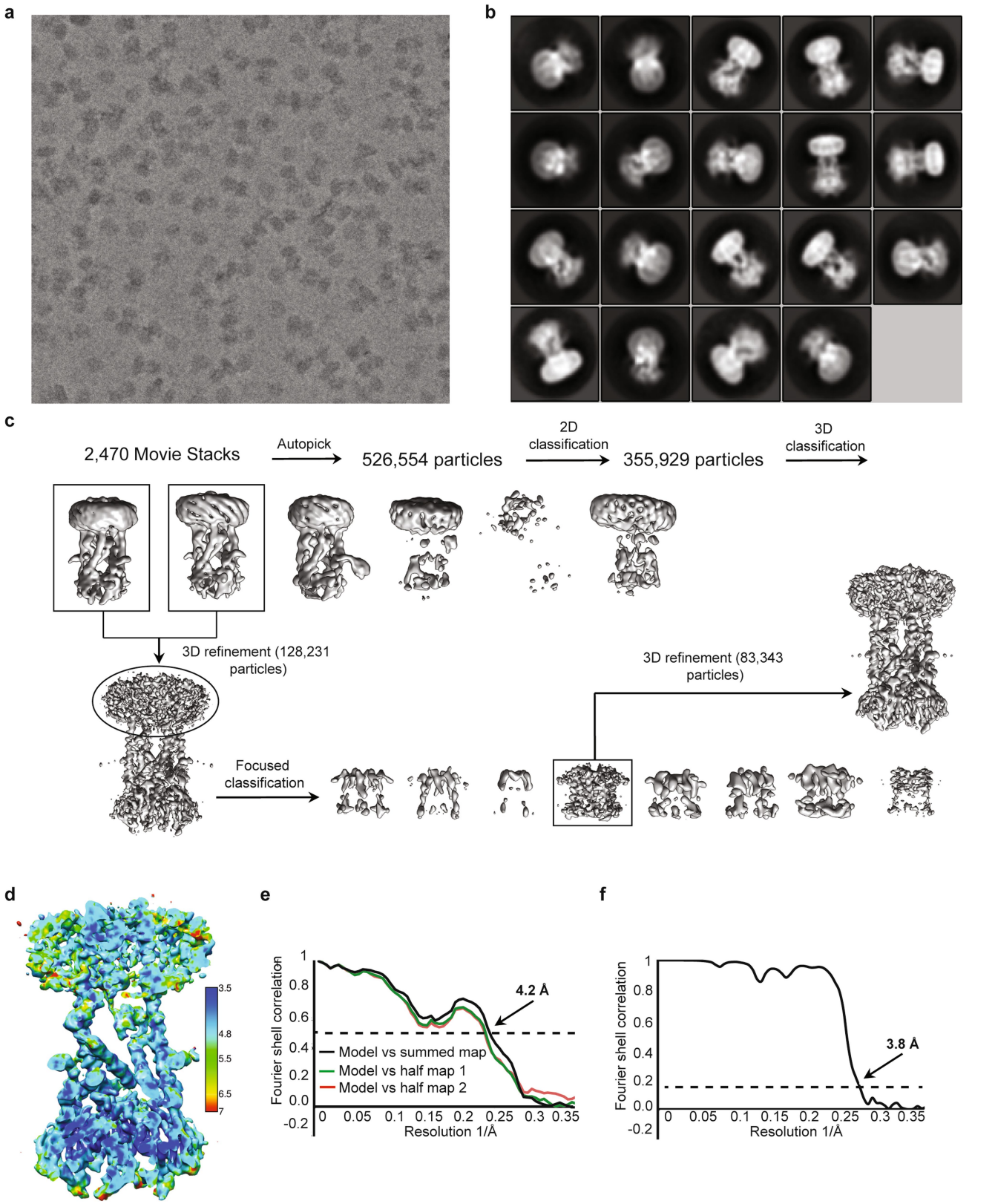
Extended Data Fig. 1 | Biochemical analysis of NfMCU. a, Representative gel filtration profile of full-length (black trace) versus truncated NfMCU used for cryo-EM studies (blue trace). **b**, Functional activity of full-length (FL) and truncated NfMCU (EM); data shown represent mean ± s.e.m. ($n = 3$ independent experiments). Cell lysates were analysed by immunoblotting using anti-His antibody to detect expression of NfMCU; RpoA was used as loading control. **c**, Detergent-solubilized NfMCU (control) separates as a single band at ~46 kDa on Coomassie-stained SDS–PAGE and crosslinks as tetramers when treated with chemical crosslinkers of varying spacer arm length. DSG, disuccinimidyl glutarate; DSS, disuccinimidyl suberate; BS_3, bis(sulfosuccinimidyl) suberate; $BSPEG_5$, PEGylated bis(sulfosuccinimidyl)suberate. See Supplementary Fig. 1 for gel source data. The experiments shown in **a** and **b** were repeated at least three times independently with similar results. The crosslinking experiment shown in **c** was performed once.

© 2018 Springer Nature Limited. All rights reserved.



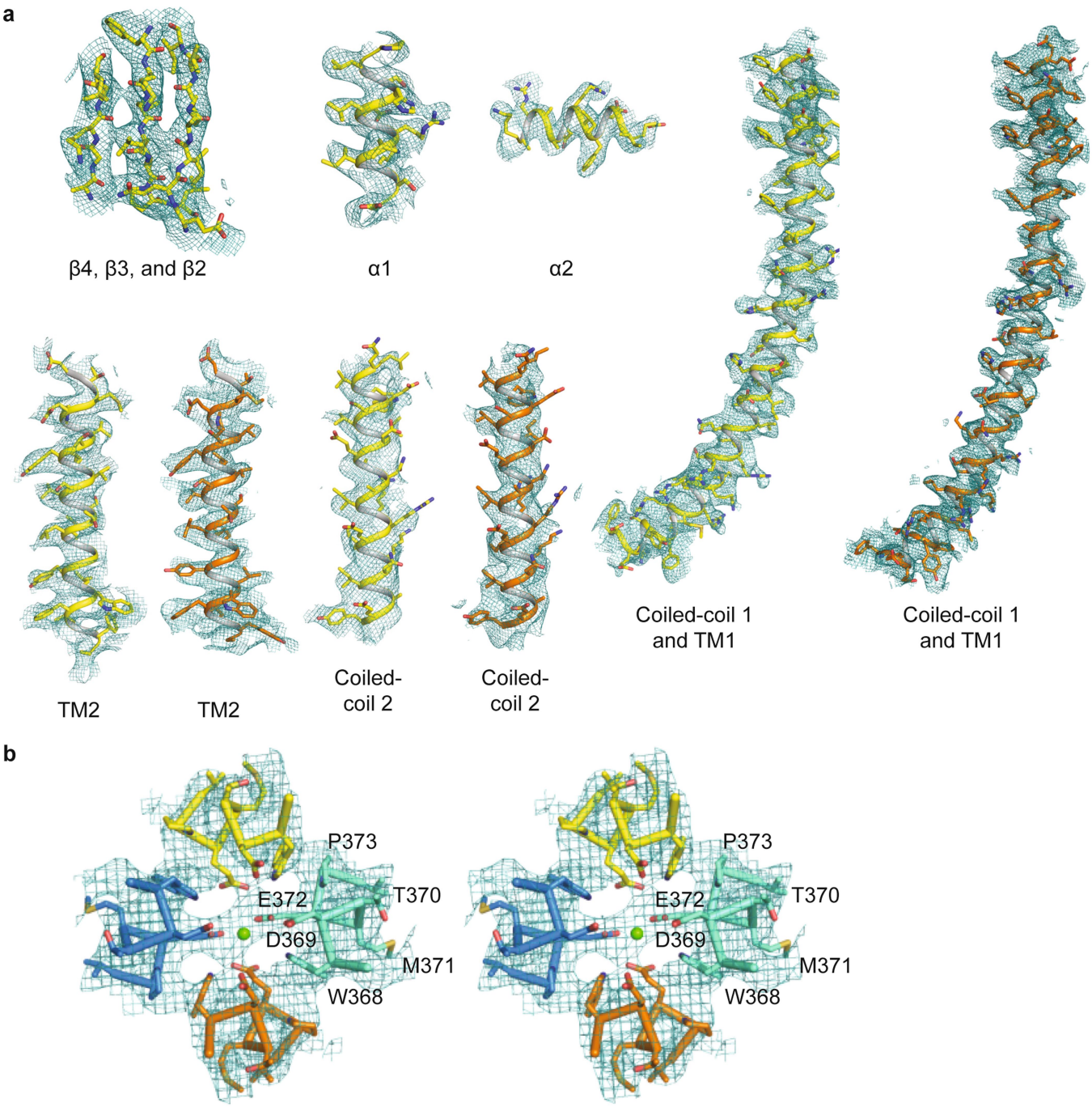
Extended Data Fig. 2 | Functional characterization of NfMCU. **a**, Calcium uptake in *E. coli* expressing empty vector (black trace) or NfMCU (blue trace), demonstrating that heterologously expressed NfMCU can confer uniporter function to bacteria. Arrows indicate addition of 20 μM $CaCl_2$ to the reaction solution. **b**, Concentration-dependent calcium uptake in *E. coli* expressing empty vector (black trace) or NfMCU (blue trace). Each point represents the rate of calcium uptake normalized to the uptake rate at 90 μM $CaCl_2$. **c**, Concentration-dependent inhibition of NfMCU by $GdCl_3$ in *E. coli* expressing NfMCU. Each point represents the rate of calcium uptake normalized to the uptake rate without $GdCl_3$ in the reaction. **d**, Concentration-dependent inhibition of Ca^{2+} uptake in *E. coli* expressing NfMCU by the MCU-specific inhibitor Ru360. Each point represents the rate of calcium uptake normalized to the uptake rate without Ru360 in the reaction. The Ru360 concentration required to reduce the Ca^{2+} uptake rate by half (IC_{50}) is calculated to be 0.5 μM. **e**, In light of the contradiction of the oligomeric state between NfMCU and cMCU-ΔNTD, we also designed an in vitro functional assay by reconstituting our purified tetrameric channel into liposomes. Unable to obtain currents of NfMCU by electrophysiology—probably owing to the channel's small conductance—we performed a radioactive ^{45}Ca flux assay on proteoliposomes loaded with 150 mM KCl. Shown here is the time-dependent $^{45}Ca^{2+}$ uptake by empty liposomes or NfMCU proteoliposomes incubated with $^{45}Ca^{2+}$ in a reaction buffer containing 150 mM KCl (red points) or 150 mM NMDG (blue points). The result demonstrates that in a reaction solution containing 150 mM NMDG and the K^+-selective ionophore valinomycin (to generate an electrical driving force), NfMCU proteoliposomes exhibited time-dependent $^{45}Ca^{2+}$ uptake while the same proteoliposomes in a reaction solution containing valinomycin and 150 mM KCl (to eliminate electrical driving force) accumulated $^{45}Ca^{2+}$ only slightly better than empty liposomes, recapitulating the voltage-dependent Ca^{2+} uptake property of the MCU. **f**, Concentration-dependent inhibition of NfMCU proteoliposomes by Ru360, demonstrating that Ru360 also blocks NfMCU $^{45}Ca^{2+}$ uptake in a concentration-dependent manner up to 50%, which is expected given an equal distribution of two channel orientations in the liposomes. Each point represents radioactivity measured after 30 min reaction normalized to sample without Ru360. All data in **a**–**f** are shown as mean ± s.e.m. ($n = 3$ independent experiments).

© 2018 Springer Nature Limited. All rights reserved.



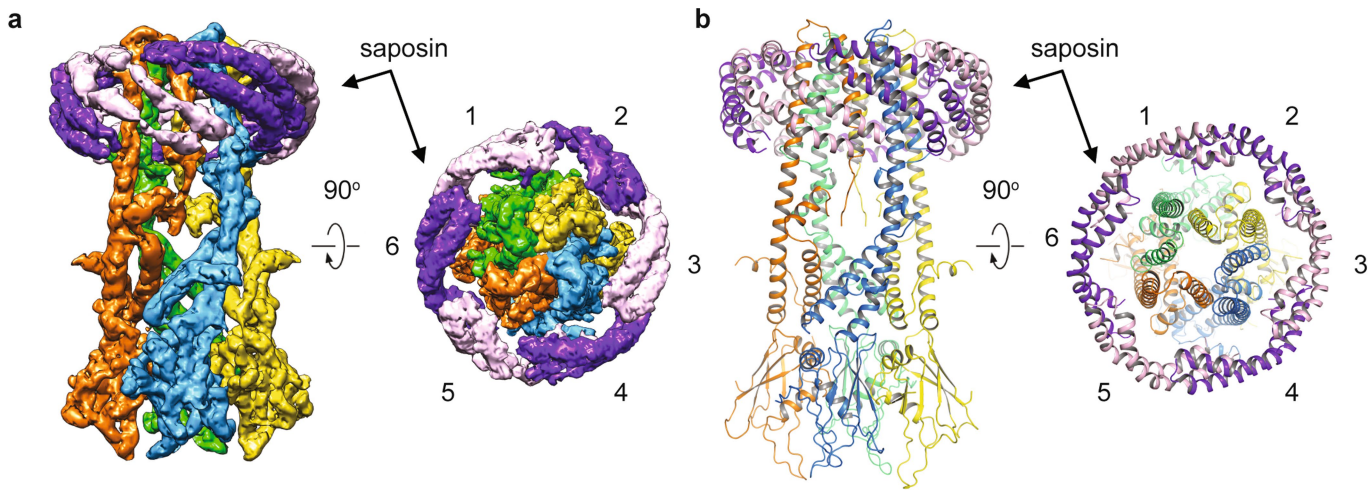
Extended Data Fig. 3 | Cryo-EM analysis of NfMCU. a, Representative phase-plate electron micrograph of the NfMCU saposin complex and 2,470 micrographs were used for structure determination. **b**, Representative 2D class averages of the NfMCU saposin complex. **c**, Flowchart of data processing to obtain the 3.8 Å cryo-EM map of the NfMCU saposin complex. Details can be found in the Image processing section of the Methods. **d**, Local resolution of NfMCU estimated with RELION2.0. **e**, FSC curves for the refined model versus the summed 3.8 Å map (black curve), the refined model versus half map 1 (red curve), and the refined model versus half map 2 that was not used for refinement (green curve). **f**, The gold-standard FSC curve for the cryo-EM map.

© 2018 Springer Nature Limited. All rights reserved.



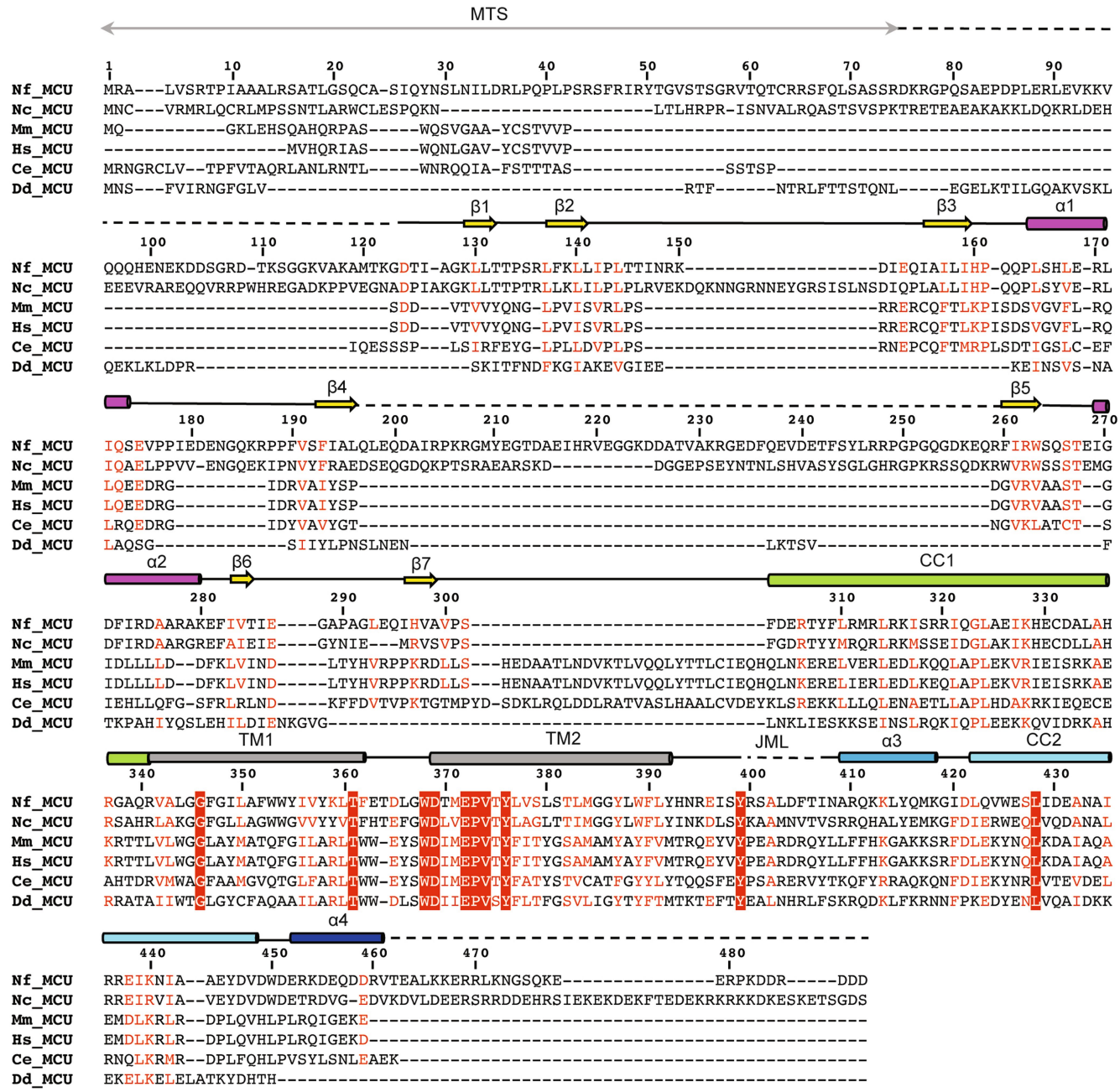
Extended Data Fig. 4 | EM map of NfMCU. a, Representative regions of the EM map of NfMCU highlighting key structural features at the NTD (α1, α2, β2– β3), the coiled-coil domain, and transmembrane domain (TM1 and TM2). Protomers 1 and 2 are shown in yellow and orange, respectively. **b**, Stereo view of the EM map carved around the TM2 filter region of NfMCU.

© 2018 Springer Nature Limited. All rights reserved.



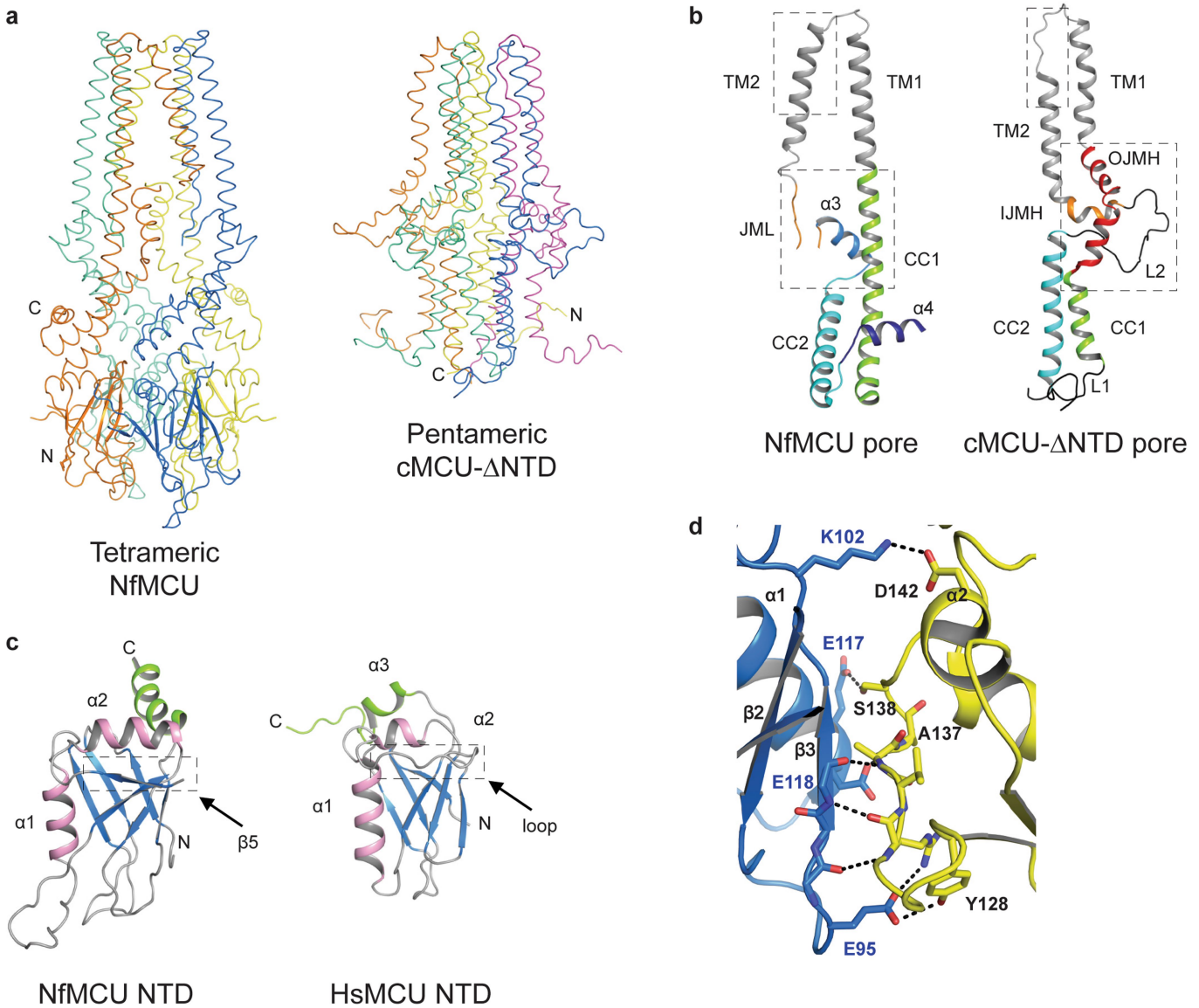
Extended Data Fig. 5 | Cryo-EM structure of NfMCU in complex with saposin. a, Cryo-EM map of NfMCU in complex with saposin (shown alternating between purple and pink). The map is low-pass filtered to 5 Å to allow better visualization of the density from the saposin molecules. Six saposin molecules, each oriented at a 45° angle to the vertical axis, wrap around NfMCU stabilized in lipid. **b**, Ribbon diagram of NfMCU in complex with saposin. Approximately 79–81 residues of the 81-residue saposin molecule were modelled as a poly-alanine chain into the helix–turn–helix EM map density for saposin.

© 2018 Springer Nature Limited. All rights reserved.



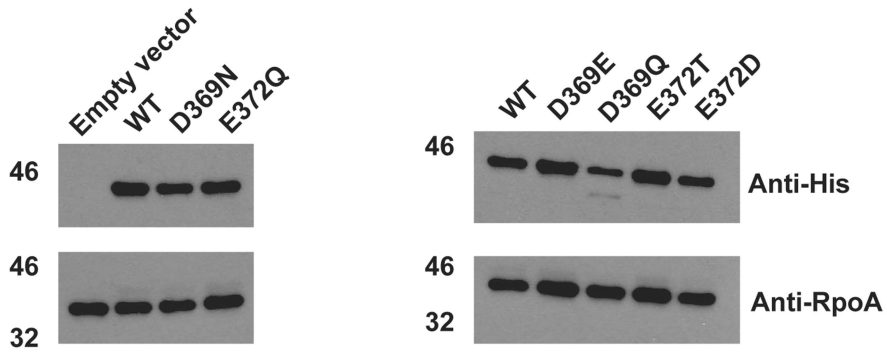
Extended Data Fig. 6 | Sequence alignment of MCU orthologues. The sequences were aligned using PROMALS3D[50] and numbered according to NfMCU. Secondary structures shown above the sequences are based on the cryo-EM structure of NfMCU. Loops are denoted by solid black lines; disordered regions not observed in the cryo-EM structure and truncations made to NfMCU are indicated by dashed lines. The first 75 residues in NfMCU (grey arrow) are predicted to be the mitochondrial targeting sequence (MTS). NCBI accession number for the sequences are: HsMCU (human MCU): NP_612366.1, MmMCU (mouse MCU): XP_006513531.1, DdMCU (*D. discoideum* MCU): XP_637750.1, CeMCU (*C. elegans* MCU): NP_500892.1, NcMCU (*Neurospora crassa* MCU): XP_959658.1, NfMCU (*Neosartorya fischeri* MCU): XP_001266985.1.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 7 | Structural comparison between MCU domains. **a**, Overall structure comparison between tetrameric NfMCU and pentameric cMCU-ΔNTD (PDB: 5ID3). Each subunit is individually coloured. **b**, Structure comparison between a single subunit of the pore domain of NfMCU and cMCU-ΔNTD. Dashed boxes indicate major differences between the two structures. Structural features in cMCU-ΔNTD are labelled as in the original structure[16]. IJMH, inner juxtamembrane helix; OJMH, outer juxtamembrane helix; L1, loop 1; L2, loop 2; JML, juxtamembrane loop; CC1, coiled-coil 1; CC2, coiled-coil 2. **c**, Structural comparison between the NTDs of NfMCU and HsMCU (PDB: 4XTB). Dashed box indicates the major difference between the two structures. **d**, Atomic interactions between two NTD subunits in the crystal structure of HsMCU NTD, equivalent to interface I identified in the cryo-EM structure of NfMCU.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 8 | Western blot analysis of NfMCU expression in *E. coli*. NfMCU, with its mitochondrial targeting sequence removed and replaced with a 6×His tag, was expressed in *E. coli* for functional analysis. Mouse anti-His antibodies (Qiagen) were used to detect NfMCU expression and mouse anti-RpoA targeting *E. coli* RNA polymerase alpha (Santa Cruz Biotechnology) was used for loading control. See Supplementary Fig. 1 for gel source data. The experiment was repeated three times independently with similar results.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Cryo-EM data collection and model statistics

	NfMCU (EMDB-7826) (PDB 6D7W)	NfMCU saposin complex (EMDB-7828) (PDB 6D80)
Data collection and processing		
Magnification	59,523	59,523
Voltage (kV)	300	300
Electron exposure (e–/Å^2)	60	60
Defocus range (μm)	-0.5	-0.5
Pixel size (Å)	0.84	0.84
Symmetry imposed	C2	C2
Initial particle images (no.)	526,554	526,554
Final particle images (no.)	128,231	128,231
Map resolution (Å)	3.8	5.0
FSC threshold	0.143	0.143
Map resolution range (Å)	184.8 – 3.8	184.8 – 5.0
Refinement		
Initial model used (PDB code)	Initial model generated in RELION2.0	Initial model generated in RELION2.0
Model resolution (Å)	3.8	5.0
FSC threshold	0.143	0.143
Model resolution range (Å)	184.8 – 3.8	184.8 – 5.0
Map sharpening *B* factor (Å^2)	-120	-120
Model composition		
Non-hydrogen atoms	8,946	11,351
Protein residues	1,087	1,568
Ligands	1	
B factors (Å^2)	103.55	135.43
Protein	103.55	135.43
Ligand	127.22	127.22
R.m.s. deviations		
Bond lengths (Å)	0.008	0.007
Bond angles (°)	1.396	1.306
Validation		
MolProbity score	2.41	2.33
Clashscore	14	13
Poor rotamers (%)	0.9	0.9
Ramachandran plot		
Favored (%)	93	93
Allowed (%)	7	7
Disallowed (%)	0	0

© 2018 Springer Nature Limited. All rights reserved.

ARTICLE

https://doi.org/10.1038/s41586-018-0330-9

X-ray and cryo-EM structures of the mitochondrial calcium uniporter

Chao Fan[1,6], Minrui Fan[1,6], Benjamin J. Orlando[2,6], Nathan M. Fastman[1,3,6], Jinru Zhang[1], Yan Xu[1], Melissa G. Chambers[2], Xiaofang Xu[1,5], Kay Perry[4], Maofu Liao[2]* & Liang Feng[1,3]*

Mitochondrial calcium uptake is critical for regulating ATP production, intracellular calcium signalling, and cell death. This uptake is mediated by a highly selective calcium channel called the mitochondrial calcium uniporter (MCU). Here, we determined the structures of the pore-forming MCU proteins from two fungi by X-ray crystallography and single-particle cryo-electron microscopy. The stoichiometry, overall architecture, and individual subunit structure differed markedly from those described in the recent nuclear magnetic resonance structure of *Caenorhabditis elegans* MCU. We observed a dimer-of-dimer architecture across species and chemical environments, which was corroborated by biochemical experiments. Structural analyses and functional characterization uncovered the roles of key residues in the pore. These results reveal a new ion channel architecture, provide insights into calcium coordination, selectivity and conduction, and establish a structural framework for understanding the mechanism of mitochondrial calcium uniporter function.

Mitochondrial Ca^{2+} uptake is pivotal for fundamental cellular processes such as ATP production, cell death, and cytosolic Ca^{2+} signalling[1–3]. Rapid mitochondrial Ca^{2+} uptake is mediated by an ion channel called the mitochondrial Ca^{2+} uniporter[4–6]. The uniporter resides in the inner mitochondrial membrane, is remarkably selective for Ca^{2+}, has high conductance, and is inhibited by ruthenium red[1,7]. During cytoplasmic Ca^{2+} signalling, local increases in the Ca^{2+} concentration activate the uniporter[8,9], leading to substantial Ca^{2+} influx.

Recent studies identified and established MCU as the essential pore-forming subunit of the Ca^{2+} conduction channel[10–13]. MCU shows higher-order oligomerization[13,14] and bears no discernible sequence similarity to other known channels. It forms a multi-component complex with regulatory subunits, including MICU[15–20], EMRE[13,21,22], MCUb[23], and MCUR1[16,24,25]. MCU is ubiquitous in plants and metazoans and present in protozoans and fungi[10,26]. By contrast, EMRE is found only in metazoans and MICU is absent from most fungi[21,26]. Malfunction or deregulation of the uniporter complex has been linked to tumour progression, muscle weakness, neurological defects, and impairments in tissue regeneration[27].

MCU is characterized by a highly conserved signature DXXE motif, two transmembrane helices, and flanking coiled-coils[26]. Its N-terminal domain (NTD) shows reliable sequence conservation only among metazoan species. This observation raises the question of whether MCU has a unified overall architecture. Recent nuclear magnetic resonance (NMR) studies of a *Caenorhabditis elegans* MCU (cMCU) core domain provided a first glimpse into the structure of MCU[28]. These investigations defined a pentamer architecture with a layered domain organization and revealed the formation of the inner core by transmembrane helix 2 (TM2) and the pore-facing location of a glutamate residue that is critical for Ca^{2+} uptake[28]. However, the cMCU construct neither contained the NTD nor conducted Ca^{2+} by itself. Furthermore, without an apparent Ca^{2+} permeation route or defined chemistry of the selectivity filter in the NMR structure, the mechanism of Ca^{2+} coordination and permeation remains unclear.

Here, we determined the structure of MCU with all structural domains, capable of conducting Ca^{2+}. Unexpectedly, our X-ray crystallography and cryo-electron microscopy (cryo-EM) studies revealed a dimer-of-dimer assembly of MCU. Both the overall architecture and the structure of each protomer were distinct from those of the NMR structure of cMCU, giving rise to a completely different pore. Our structures define the key elements of the Ca^{2+} permeation pathway and provide insights into Ca^{2+} selectivity.

Functional characterization and crystal structure

We identified two MCUs from fungi, *Fusarium graminearum* MCU (FgMCU) and *Metarhizium acridum* MCU (MaMCU) (Extended Data Fig. 1a), that showed excellent biochemical behaviour (Extended Data Fig. 1b, c). As EMRE is absent in fungi[21] and another fungal MCU was shown to mediate mitochondrial Ca^{2+} uptake[29], these MCUs were potentially sufficient for basic uniporter activity. In an *Escherichia coli*-based system with the fluorescent Ca^{2+} reporter GCaMP2[30], cells with wild-type MaMCU (Fig. 1a) or FgMCU (Extended Data Fig. 1d) showed robust Ca^{2+} uptake. This uptake was diminished by mutation of residues critical for Ca^{2+} conductance[10,11] or addition of the classic uniporter inhibitor ruthenium red (Fig. 1a, Extended Data Fig. 1d). Furthermore, when reconstituted in liposomes, purified wild-type MaMCU (Fig. 1b) and FgMCU (Extended Data Fig. 1e) showed robust ion conduction activity in a fluorescence-based flux assay[31]. The addition of Gd^{3+}, another MCU inhibitor, abolished this activity (Fig. 1b, Extended Data Fig. 1e). Thus, MaMCU and FgMCU recapitulate key features of MCU and MCU is sufficient to form a pore.

We obtained diffracting crystals of MaMCU either by inserting rubredoxin into an NTD loop or with nanobody bound (Extended Data Fig. 2a, b, d and Extended Data Table 1). As crystal with nanobody diffracted X-rays better (to approximately 3.1 Å) and showed considerably better electron density in the transmembrane region, it was used for all structural interpretations. With 12 mercury sites introduced individually into the transmembrane region (Extended Data Fig. 2e) and clear

[1]Department of Molecular and Cellular Physiology, Stanford University School of Medicine, Stanford, CA, USA. [2]Department of Cell Biology, Harvard Medical School, Boston, MA, USA. [3]Biophysics Program, Stanford University, Stanford, CA, USA. [4]NE-CAT and Dept. of Chemistry and Chemical Biology, Cornell University, Argonne National Laboratory, Argonne, IL, USA. [5]Present address: Department of Ophthalmology, Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai Key Laboratory of Orbital Diseases and Ocular Oncology, Shanghai, China. [6]These authors contributed equally: Chao Fan, Minrui Fan, Benjamin J. Orlando, Nathan M. Fastman. *e-mail: maofu_liao@hms.harvard.edu; liangf@stanford.edu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

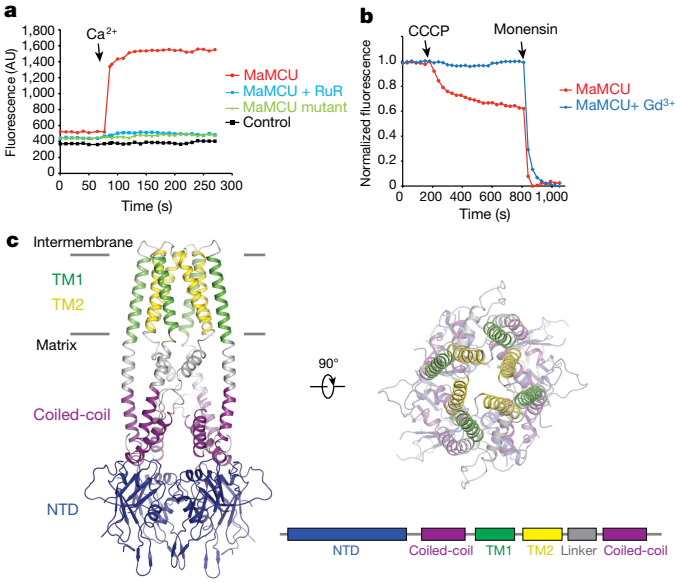


Fig. 1 | Functional characterization and crystal structure of MCU. **a**, Ca^{2+} uptake activity of MaMCU in *E. coli*. Fluorescence changes upon addition of 0.5 mM Ca^{2+} are shown for wild-type (red), D333A/E336A (green), and control (black). Ruthenium red (RuR) abolished wild-type uptake (blue). AU, arbitrary units. **b**, Cation flux activity of MaMCU in reconstituted liposomes. MCU efficiently conducts monovalent cations (for example, Na^+) when divalent cations are absent[6]. Vesicles containing 150 mM Na^+ were assayed in buffer containing 150 mM *N*-methyl-D-glucamine (NMDG). Upon addition of proton ionophore (CCCP), Na^+ efflux through MCU drives CCCP-mediated proton influx, quenching 9-amino-6-chloro-2-methoxyacridine (ACMA) fluorescence. Liposomes with MaMCU showed robust fluorescence changes (red); Gd^{3+} addition abolished flux (blue). Na^+ ionophore (monensin) was added to make liposomes permeable to Na^+. **c**, MaMCU crystal structure. Ribbon representations of MaMCU viewed from membrane (left) or cellular side (right). Each domain is coloured according to subunit diagram (bottom right). Data in **a** and **b** are representative of three independent experiments with similar results.

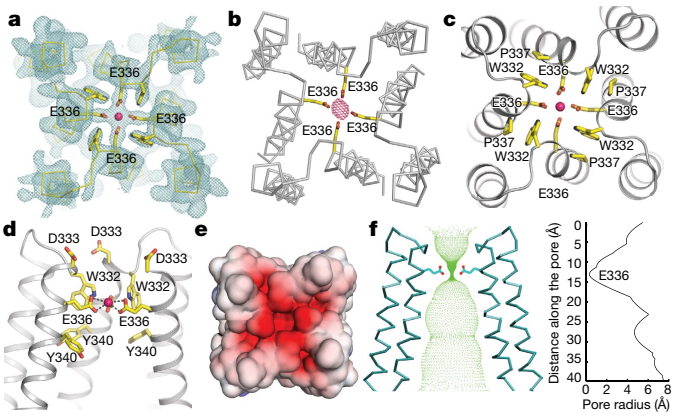


Fig. 2 | Ca^{2+} permeation pathway. **a**, Electron density map around ion coordination region. Weighted $2F_o-F_c$ contoured at 3σ (cyan) and Ca^{2+} ion (pink sphere) are shown. **b**, Anomalous-difference Fourier map for Ca^{2+} at 4.5 Å contoured at 4σ (pink mesh). **c**, Ca^{2+} coordination in filter region viewed from top. Refined model (ribbon) and selected residues (sticks) are shown. **d**, Ca^{2+} coordination in filter region viewed within membrane, with front helices removed for clarity. **e**, Surface representation viewed from top, coloured according to electrostatic potential (red, negative; blue, positive). **f**, Ion permeation pathway. Ion-conducting passage (green dots) was calculated by HOLE. Pore radius is shown on right.

side chain densities (Extended Data Fig. 2f, g), we could unambiguously register and build the side chains.

In the crystal, two crystallographically related dimers interact tightly, forming one MCU channel with a two-fold symmetry axis along the Ca^{2+}-conducting pathway. MCU comprises three layers: a transmembrane domain, a coiled-coil, and an NTD, forming an elongated shape (Fig. 1c). Notably, the oligomerization state, architecture, and single-subunit structure (including the selectivity filter configuration, packing mode between transmembrane helices, and overall fold) of MaMCU differed from the published pentameric structure of cMCU (Extended Data Fig. 3a, b).

Transmembrane domain and ion conduction pore

The transmembrane domain has an approximately four-fold symmetry and the shape of a truncated pyramid (Fig. 1c). The two transmembrane helices (TM1 and TM2) connect by a short loop. TM1 is located on the outside and TM2 lines the central ion conduction pore, consistent with the observation that TM1 of human MCU mediates the interaction with EMRE[22]. TM1 interacts loosely with TM2 of the same subunit and packs intimately against TM2 from a neighbouring subunit (Fig. 1c), producing a sense of domain swapping in the transmembrane connectivity. In one pair of subunits, the connection between TM2 and the second helix of the coiled-coil is not well ordered, indicating flexibility.

The pore of MCU is narrow at the top. Notably, in contrast to the assumption that the signature motif DXXE lies within a loop, DXXE spans the first helical turn of TM2 with the strictly conserved Asp333 and Glu336 constituting the first two pore-lining residues (Fig. 2d). Asp333 is located at the intermembrane end of TM2 and is solvent exposed. Its side chain density was less resolved than those of other key residues. In the modelled configuration, the Asp333 side chain carboxylate does not directly point to the pore centre, potentially providing negative charge for Ca^{2+} recruitment (Fig. 2e). Alternatively, if Asp333 takes the most common rotamer position, which points into the pore, its carboxylate ring could coordinate a hydrated Ca^{2+}, given a diameter of approximately 8 Å. One helical turn below Asp333 is Glu336, whose carboxylate group points to the pore centre (Fig. 2a, c). We observed a strong density along the central axis, consistent with a bound ion surrounded by the carboxyl oxygens of Glu336 (Fig. 2a). As the protein was purified and crystallized in the presence of 2 mM Ca^{2+} and no other divalent ions, we assigned Ca^{2+} at this location, which was further confirmed by a strong anomalous difference peak occupying the same location (Fig. 2b). The diameter of the Glu336 carboxylate ring is approximately 5 Å, too small for hydrated Ca^{2+} to go through. The observed Ca^{2+}–oxygen distance (around 2.4–2.7 Å; Fig. 2d) is consistent with direct coordination[32]. We thus propose that Ca^{2+} passes through a selectivity filter formed by the glutamate ring in a dehydrated form, similar to the direct coordination of Ca^{2+} by four aspartate side chains in TRPV6[33]. Between the side chains of two neighbouring Glu336 is the indole ring of Trp332, which stabilizes the closely spaced carboxyl side chains of Glu336 through hydrogen bonds or anion-π interactions (Fig. 2c). Trp332 also forms a close stacking interaction with Pro337 that is important for orientating Glu336 for Ca^{2+} coordination. The intimately interacting Trp332, Glu336, and Pro337 are thus essential for proper configuration of the selectivity filter, consistent with their strict conservation across species.

One helical turn below Glu336 (and two below Asp333) is Tyr340. The aromatic side chain of Tyr340 is lined along the wall of the pore, forming a passage about 12 Å wide. Below the tyrosine, the pore is relatively wide open (Fig. 2f), consistent with the high conductance of MCU. Near the matrix side, TM1 and TM2 of the same subunit make fewer direct contacts. Perhaps lipids fill in to seal the bottom of the channel, in line with weak extra density around this location.

Structure of the soluble domain

Within the mitochondrial matrix resides the soluble domain of MCU, comprised of an NTD, a coiled-coil, and a variable non-conserved C terminus. This domain is involved in MCU assembly, regulation of

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

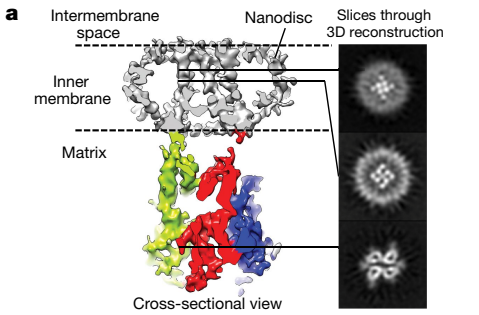


Fig. 3 | Cryo-EM structure of FgMCU in nanodiscs. a, Slice view of the FgMCU 3D reconstruction (type 1 conformation) filtered to 5 Å resolution. Nanodisc and transmembrane helices (grey) and soluble domains (red, green, cyan, and blue) are shown. Slice views are shown at indicated levels. **b**, Overlay of FgMCU cryo-EM map with MaMCU structure model. Protomers are shown in different colours. Data collection and map calculation were not repeated.

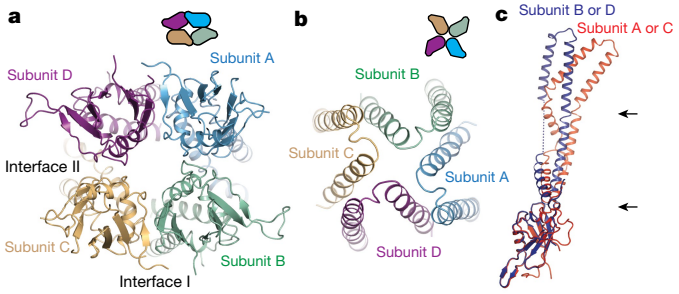


Fig. 4 | Symmetry mismatch in domain organization. a, Matrix-side view of MaMCU NTD. Domain organization is depicted using schematic drawings. **b**, Cellular-side view of MaMCU transmembrane domain. **c**, Structural comparison of subunits (superimposed through the NTD). Arrows indicate potential hinge regions.

channel activity, and post-translational modification[14,25,34]. We determined the crystal structure of the soluble domain of MaMCU at 3.1 Å resolution (Extended Data Fig. 3c), providing further support of its quaternary organization.

Notably, the globular NTD of MaMCU has the same structural fold as that of human MCU (root mean square deviation (r.m.s.d.) = 1.6 Å) despite there being no significant sequence similarity (Extended Data Fig. 3d). Within the crystal lattice, the soluble domain forms a dimer-of-dimers configuration (Extended Data Fig. 2c), matching well with the corresponding part of the whole MCU structure. Notably, a dimer interface of NTDs in MaMCU is the same as an interface of human NTD found in multiple crystal forms[14,35] (Extended Data Fig. 3e). These observations support a conserved architecture from fungi to humans.

Single-particle electron microscopy of MCU

We used negative-stain electron microscopy to screen FgMCU and MaMCU exchanged into amphipols (A8–35 and PMAL-C8) or reconstituted into lipid nanodiscs. The negatively stained MCU proteins were monodisperse and showed two elongated densities extending parallel from the membrane region (Extended Data Fig. 4a). Cryo-EM analysis was carried out on MaMCU in A8–35, FgMCU in PMAL-C8, and FgMCU in nanodiscs (Extended Data Figs. 4, 5). The 2D averages showed the soluble domains organized with an approximate two-fold symmetry. Following 3D classification without symmetry (Extended Data Figs. 4, 5), the 3D reconstructions consistently showed a dimer-of-dimers structure. 3D refinement with two-fold symmetry generated final maps at 5–7 Å resolution (Extended Data Table 2). The overall architectures of all three MCU complexes were essentially identical.

In the best cryo-EM 3D reconstructions using FgMCU in nanodiscs (Extended Data Fig. 5), all eight transmembrane helices were well resolved and arranged with a near four-fold symmetry (Fig. 3a). The resolution was likely to be limited by the mobile transmembrane helices and flexible connection between the soluble and transmembrane domains. Focused 3D classification revealed different structures (Extended Data Fig. 5c). One major class represented the conformation in the averaged cryo-EM map (conformation type 1; 19% of particles); another generated a 3D reconstruction showing less complete transmembrane helices but strong central densities, presumably contributed by hydrated ions or small molecules (conformation type 2; 21% of particles). These two classes were refined to 5 Å and 4.8 Å resolution, respectively, for the transmembrane domain (Extended Data Fig. 6). The type 1 conformation superimposes well onto the structure model of MaMCU (Fig. 3b and Extended Data Fig. 6g). In the type 2 conformation, the lower halves of two opposing inner transmembrane helices appear to bend towards the centre (Extended Data Fig. 6g), possibly representing a more closed conformation.

Stoichiometry and symmetry mismatch of MCU

To test whether our observed tetrameric organization represents the quaternary structure of MCU in the membrane, we performed site-directed disulfide bridge crosslinking. On the basis of the structure, we introduced a pair of cysteines at the protomer interface of the transmembrane domain on a cysteine-free background (Extended Data Fig. 7a). In either detergent micelle or the isolated membrane, crosslinking under oxidizing conditions gave rise to a newly formed product with the expected size of a tetramer in a time-dependent manner (Extended Data Fig. 7b, c). Furthermore, MaMCU crosslinked in the membrane migrated in the same way as untreated protein on size-exclusion chromatography (Extended Data Fig. 7d). To evaluate whether the architecture of MCU is conserved in metazoan MCU, we made equivalent double cysteine mutations in MCU from mosquito, which possesses EMRE and MICU in the genome. Under oxidation conditions, the main cross-linked product migrated to the same tetramer position as that of MaMCU on SDS–PAGE (Extended Data Fig. 7e). Comparison of 2D averages of negatively stained mosquito MCU with those of FgMCU indicated a similar overall architecture (Extended Data Fig. 7f).

Overall, MCU is organized as a dimer of dimers. The soluble domain shows two-fold symmetry, with each NTD directly contacting two neighbouring subunits through distinct interfaces (Fig. 4a). By contrast, the transmembrane domain shows approximately four-fold rotational symmetry (Fig. 4b). This symmetry mismatch is evident from the slice views of the cryo-EM map (Fig. 3a) and is illustrated by markers on each domain in the crystal structure (Extended Data Fig. 7g). Superimposing subunits A and C with neighbouring subunits B and D revealed how the symmetry mismatch might be reconciled (Fig. 4c). First, the coiled-coil domain swings approximately 8° around its end near the NTD. This movement propagates through a long distance to the top of the transmembrane domain (about 85 Å away), resulting in a large displacement of the transmembrane domain. Second, the junction between the transmembrane domain and the coiled-coil is flexible and accommodates a roughly 20° bend right after the end of the coiled-coil, moving the transmembrane domain farther away. As a result, subunits A and C adopt a substantially different conformation from subunits B and D, even though each individual domain superimposes well. Both ends of the coiled-coil are likely to function as joints, providing flexibility for different relative orientations.

Functional studies of transmembrane domain

To validate the structural model and assess the functional role of residues in the transmembrane domain, we performed a systematic alanine scan. Consistent with the structural observations, single point mutations that substantially impaired uptake activity were mostly concentrated on the pore-facing side of TM2, with a helical periodic distribution pattern (Fig. 5a, b). The strictly conserved Trp332, Pro337, and Glu336, which are essential for proper configuration of the selectivity filter, were indispensable for uptake. Substitution of Tyr340, which sits

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

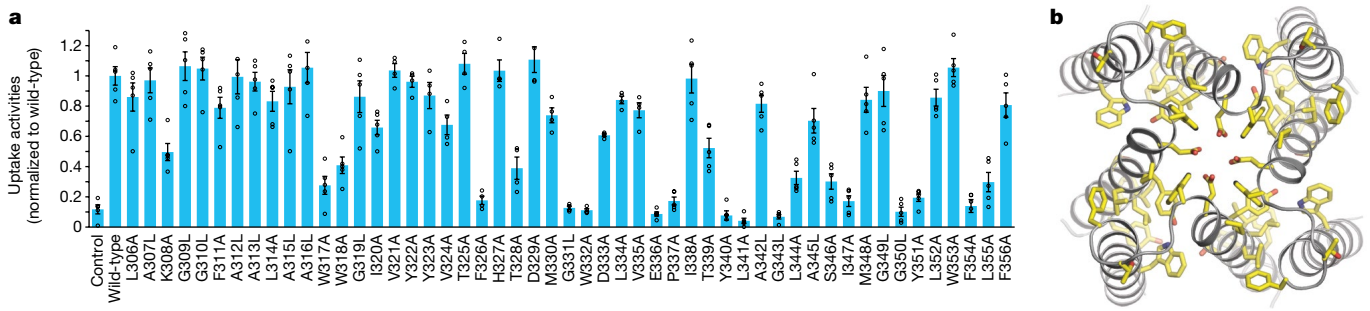


Fig. 5 | Functional validation of transmembrane domain. a, *E. coli*-based Ca^{2+} uptake activities of MaMCU with individual alanine (or leucine for alanine/glycine) point mutations in transmembrane domain in the presence of 50 μM Ca^{2+} (mean ± s.e.m., $n = 5$ independent experiments except for mutants on residues 321–335, $n = 4$ independent experiments). **b**, Function-impairing residues. Transmembrane helices (grey) and residues whose substitution substantially affected Ca^{2+} uptake (below arbitrary 0.4 cut-off; yellow sticks) are shown.

directly below the filter, substantially impaired uptake and proper protein folding, indicating an important structural role while complicating specific functional interpretation. The residue on TM1 that showed the most pronounced effect on uptake activity was Trp317, whose side chain constitutes a primary contact point between TM1 and TM2. Mutating Phe326 or Gly331 of the TM1–TM2 linker also impaired Ca^{2+} uptake, probably owing to their effect on the linker conformation and thus the configuration of the pore entrance.

We examined the effect of the mutations on the sensitivity of MCU to ruthenium red by testing residues near the intermembrane surface. Substitution of D333A had the most marked effect and substantially reduced sensitivity to ruthenium red (Extended Data Fig. 8), consistent with an equivalent mutation of the human MCU[36]. Ser259 of the human MCU, another residue involved in ruthenium red sensitivity, is not conserved in MaMCU or FgMCU[10,11]. Perhaps positively charged ruthenium red interacts with negatively charged Asp333 near the entrance of the pore and blocks the channel, competing with the Ca^{2+}-binding site. This would explain why ruthenium red inhibited inward but not outward current of MCU[6]. Together, our mutagenesis studies are fully compatible with the structure, providing insights into MCU function.

Discussion

Our X-ray and electron microscopy structures, together with membrane cross-linking and functional experiments, have revealed a conserved dimer-of-dimers architecture of MCU with physiological relevance. The marked difference between our MCU and the cMCU structures might be related to the approaches used, as the cMCU sample preparation involved the relatively harsh detergent Fos-choline-14 and inclusion bodies, and the oligomerization state of cMCU was inferred from an 18 Å-resolution negative-stain electron microscopy reconstruction[28].

The channel domain of MCU represents a previously undescribed architecture of ion channels. Compared with voltage-gated ion channels (VGICs), MCU has a roughly inverted shape resembling a truncated pyramid. Moreover, the selectivity filter of MCU is formed by side chains on TM2, as opposed to a reentry loop lining the pore for VGICs[37–44]. Nonetheless, MCU, Ca^{2+}-conducting VGICs, and the hexameric Orai Ca^{2+} channel share a common theme: a carboxylate ring at the entrance of the selectivity filter is likely to provide negative charge to effectively coordinate Ca^{2+}. This convergent configuration indicates a fundamental requirement for Ca^{2+} channels that is achieved by different means. Compared to Orai, the pore of MCU is mostly hydrophilic and wider open, which may contribute to its relatively high conductance.

MCU has high selectivity with a high-affinity binding site to Ca^{2+} in the pore[6], which is essential for its physiological function, given the relatively low cytosolic Ca^{2+} concentration. We propose that the high selectivity of Ca^{2+} is achieved by high-affinity binding through direct coordination by the carboxylate oxygen of the Glu336 side chains. The conserved interaction network among Trp332, Glu336, and Pro337 stabilizes closely spaced carboxylate groups of the Glu336 ring, which might contribute to the high-affinity binding of Ca^{2+}. Asp333 and Glu336 of MCU showed a notable parallel with the engineered Ca_vAb Ca^{2+} channel, in which two adjacent rings of negatively charged residues also line the mouth of the pore and are important for Ca^{2+} selectivity[45]. We observed only one strong Ca^{2+} binding site in the crystal structure. It remains to be investigated whether Asp333 is also involved in coordinating a secondary Ca^{2+} for a potential knock-off mechanism[45].

Post-translational modifications or divalent cation binding to the NTD can modulate Ca^{2+} conduction by MCU[14,25,34]. In this regard, it is notable that the symmetry mismatch between the transmembrane domain and the soluble NTD is similar to that of ionotropic glutamate receptors (iGluRs)[40]. During the gating cycle of iGluRs, the ligand-binding domain undergoes substantial re-organization and transitions between two-fold and four-fold symmetry[46]. By analogy, a similar reorganization in the NTD could modulate MCU channel activity by switching subunits A and C or B and D between alternate conformations, thereby altering the packing or conformation of the transmembrane helices. In the crystal structure of MaMCU, immediately below the selectivity filter, Tyr340 interacts with residues on the neighbouring subunit, orienting its side chain along the pore. Notably, the cryo-EM map of FgMCU shows density in the pore that could potentially result from one pair of Tyr340 side chains pointing into the pore. Although higher resolution is needed for a definitive assignment, an intriguing possibility is that altered transmembrane packing may weaken the interactions of Tyr340 with nearby transmembrane helices to facilitate a rotamer switch of one pair of tyrosines to control Ca^{2+} flow through the pore. Alternatively, the rearrangement of NTDs could potentially lead to bending of transmembrane helices to constrict the pore, as suggested by two possible conformations in the cryo-EM map.

In summary, our structures and biochemical studies establish the stoichiometry and architecture of MCU and provide important insights into Ca^{2+} coordination and permeation. In the future, structures at higher resolutions, in various states, and in complex with regulatory components will be needed for a detailed mechanistic understanding of Ca^{2+} selectivity and channel gating.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0330-9.

Received: 7 January 2018; Accepted: 17 May 2018;
Published online 11 July 2018.

1. Gunter, T. E. & Pfeiffer, D. R. Mechanisms by which mitochondria transport calcium. *Am. J. Physiol.* **258**, C755–C786 (1990).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

2. Rizzuto, R., De Stefani, D., Raffaello, A. & Mammucari, C. Mitochondria as sensors and regulators of calcium signalling. *Nat. Rev. Mol. Cell Biol.* **13**, 566–578 (2012).
3. Szabadkai, G. & Duchen, M. R. Mitochondria: the hub of cellular Ca^{2+} signaling. *Physiology (Bethesda)* **23**, 84–94 (2008).
4. Deluca, H. F. & Engstrom, G. W. Calcium uptake by rat kidney mitochondria. *Proc. Natl Acad. Sci. USA* **47**, 1744–1750 (1961).
5. Vasington, F. D. & Murphy, J. V. Ca ion uptake by rat kidney mitochondria and its dependence on respiration and phosphorylation. *J. Biol. Chem.* **237**, 2670–2677 (1962).
6. Kirichok, Y., Krapivinsky, G. & Clapham, D. E. The mitochondrial calcium uniporter is a highly selective ion channel. *Nature* **427**, 360–364 (2004).
7. De Stefani, D., Rizzuto, R. & Pozzan, T. Enjoy the trip: calcium in mitochondria back and forth. *Annu. Rev. Biochem.* **85**, 161–192 (2016).
8. Rizzuto, R., Brini, M., Murgia, M. & Pozzan, T. Microdomains with high Ca^{2+} close to IP_3-sensitive channels that are sensed by neighboring mitochondria. *Science* **262**, 744–747 (1993).
9. Giacomello, M. et al. Ca^{2+} hot spots on the mitochondrial surface are generated by Ca^{2+} mobilization from stores, but not by activation of store-operated Ca^{2+} channels. *Mol. Cell* **38**, 280–290 (2010).
10. Baughman, J. M. et al. Integrative genomics identifies MCU as an essential component of the mitochondrial calcium uniporter. *Nature* **476**, 341–345 (2011).
11. De Stefani, D., Raffaello, A., Teardo, E., Szabò, I. & Rizzuto, R. A forty-kilodalton protein of the inner membrane is the mitochondrial calcium uniporter. *Nature* **476**, 336–340 (2011).
12. Chaudhuri, D., Sancak, Y., Mootha, V. K. & Clapham, D. E. MCU encodes the pore conducting mitochondrial calcium currents. *eLife* **2**, e00704 (2013).
13. Kovács-Bogdán, E. et al. Reconstitution of the mitochondrial calcium uniporter in yeast. *Proc. Natl Acad. Sci. USA* **111**, 8985–8990 (2014).
14. Lee, S. K. et al. Structural insights into mitochondrial calcium uniporter regulation by divalent cations. *Cell Chem. Biol.* **23**, 1157–1169 (2016).
15. Csordás, G. et al. MICU1 controls both the threshold and cooperative activation of the mitochondrial Ca^{2+} uniporter. *Cell Metab.* **17**, 976–987 (2013).
16. Mallilankaraman, K. et al. MCUR1 is an essential component of mitochondrial Ca^{2+} uptake that regulates cellular metabolism. *Nat. Cell Biol.* **14**, 1336–1343 (2012).
17. Kamer, K. J., Grabarek, Z. & Mootha, V. K. High-affinity cooperative Ca^{2+} binding by MICU1-MICU2 serves as an on-off switch for the uniporter. *EMBO Rep.* **18**, 1397–1411 (2017).
18. Perocchi, F. et al. MICU1 encodes a mitochondrial EF hand protein required for Ca^{2+} uptake. *Nature* **467**, 291–296 (2010).
19. Liu, J. C. et al. MICU1 serves as a molecular gatekeeper to prevent in vivo mitochondrial calcium overload. *Cell Reports* **16**, 1561–1573 (2016).
20. Patron, M. et al. MICU1 and MICU2 finely tune the mitochondrial Ca^{2+} uniporter by exerting opposite effects on MCU activity. *Mol. Cell* **53**, 726–737 (2014).
21. Sancak, Y. et al. EMRE is an essential component of the mitochondrial calcium uniporter complex. *Science* **342**, 1379–1382 (2013).
22. Tsai, M. F. et al. Dual functions of a small regulatory subunit in the mitochondrial calcium uniporter complex. *eLife* **5**, e15545 (2016).
23. Raffaello, A. et al. The mitochondrial calcium uniporter is a multimer that can include a dominant-negative pore-forming subunit. *EMBO J.* **32**, 2362–2376 (2013).
24. Chaudhuri, D., Artiga, D. J., Abiria, S. A. & Clapham, D. E. Mitochondrial calcium uniporter regulator 1 (MCUR1) regulates the calcium threshold for the mitochondrial permeability transition. *Proc. Natl Acad. Sci. USA* **113**, E1872–E1880 (2016).
25. Tomar, D. et al. MCUR1 Is a scaffold factor for the MCU complex function and promotes mitochondrial bioenergetics. *Cell Reports* **15**, 1673–1685 (2016).
26. Bick, A. G., Calvo, S. E. & Mootha, V. K. Evolutionary diversity of the mitochondrial calcium uniporter. *Science* **336**, 886 (2012).
27. Mammucari, C., Gherardi, G. & Rizzuto, R. Structure, activity regulation, and role of the mitochondrial calcium uniporter in health and disease. *Front. Oncol.* **7**, 139 (2017).
28. Oxenoid, K. et al. Architecture of the mitochondrial calcium uniporter. *Nature* **533**, 269–273 (2016).
29. Song, J., Liu, X., Zhai, P., Huang, J. & Lu, L. A putative mitochondrial calcium uniporter in *A. fumigatus* contributes to mitochondrial Ca^{2+} homeostasis and stress responses. *Fungal Genet. Biol.* **94**, 15–22 (2016).
30. Shigetomi, E., Kracun, S., Sofroniew, M. V. & Khakh, B. S. A genetically targeted optical sensor to monitor calcium signals in astrocyte processes. *Nat. Neurosci.* **13**, 759–766 (2010).
31. Hou, X., Pedi, L., Diver, M. M. & Long, S. B. Crystal structure of the calcium release-activated calcium channel Orai. *Science* **338**, 1308–1313 (2012).
32. Yang, W., Lee, H. W., Hellinga, H. & Yang, J. J. Structural analysis, identification, and design of calcium-binding sites in proteins. *Proteins* **47**, 344–356 (2002).
33. Saotome, K., Singh, A. K., Yelshanskaya, M. V. & Sobolevsky, A. I. Crystal structure of the epithelial calcium channel TRPV6. *Nature* **534**, 506–511 (2016).
34. Dong, Z. et al. Mitochondrial Ca^{2+} uniporter is a mitochondrial luminal redox sensor that augments MCU channel activity. *Mol. Cell* **65**, 1014–1028.e7 (2017).
35. Lee, Y. et al. Structure and function of the N-terminal domain of the human mitochondrial calcium uniporter. *EMBO Rep.* **16**, 1318–1333 (2015).
36. Arduino, D. M. et al. Systematic identification of MCU modulators by orthogonal interspecies chemical screening. *Mol. Cell* **67**, 711–723 (2017).
37. Liao, M., Cao, E., Julius, D. & Cheng, Y. Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature* **504**, 107–112 (2013).
38. Doyle, D. A. et al. The structure of the potassium channel: molecular basis of K^+ conduction and selectivity. *Science* **280**, 69–77 (1998).
39. Yan, Z. et al. Structure of the rabbit ryanodine receptor RyR1 at near-atomic resolution. *Nature* **517**, 50–55 (2015).
40. Sobolevsky, A. I., Rosconi, M. P. & Gouaux, E. X-ray structure, symmetry and mechanism of an AMPA-subtype glutamate receptor. *Nature* **462**, 745–756 (2009).
41. Fan, G. et al. Gating machinery of $InsP_3R$ channels revealed by electron cryomicroscopy. *Nature* **527**, 336–341 (2015).
42. Zalk, R. et al. Structure of a mammalian ryanodine receptor. *Nature* **517**, 44–49 (2015).
43. Efremov, R. G., Leitner, A., Aebersold, R. & Raunser, S. Architecture and conformational switch mechanism of the ryanodine receptor. *Nature* **517**, 39–43 (2015).
44. Wu, J. et al. Structure of the voltage-gated calcium channel $Ca_v1.1$ at 3.6 Å resolution. *Nature* **537**, 191–196 (2016).
45. Tang, L. et al. Structural basis for Ca^{2+} selectivity of a voltage-gated calcium channel. *Nature* **505**, 56–61 (2014).
46. Zhu, S. & Gouaux, E. Structure and symmetry inform gating principles of ionotropic glutamate receptors. *Neuropharmacology* **112**, 11–15 (2017).

Acknowledgements We thank the staff at Beamline 24ID-C/E, 23ID-B/D, and 17ID (APS, Argonne National Laboratory), 12-2 (SSRL, SLAC National Laboratory), and 5.0.2 (Advanced Light Source). We thank Z. Li at HMS and C. Xu and K. Song at UMass cryo-EM facility for help with EM data collection. This work was made possible by support from Stanford University, the Klingenstein-Simons Fellowship, and the Harold and Leila Y. Mathers Charitable Foundation to L.F., and from a National Science Foundation Graduate Research Fellowship to N.M.F. NE-CAT is supported by NIH P41 GM103403 and S10 RR029205. We thank A. Manglik for discussions and R. Lewis for suggestions and comments.

Reviewer information *Nature* thanks M. Prakriya, D. Stokes and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions C.F. and M.F. performed protein purification, biochemistry, crystallization, data collection, crystallography and functional assays. B.J.O. performed negative-stain EM, cryo-EM, and data processing. N.M.F. performed protein engineering, crystallization, and data collection. Y.X. performed protein purification and biochemistry. J.Z. performed crystallography. M.G.C. helped with initial EM screening. X.X. helped with initial characterization. K.P. assisted with crystallography. M.L. oversaw the EM studies and contributed to EM data collection and processing. L.F. conceived the project, oversaw the X-ray structural, biochemical, and functional studies and contributed to experimental work and crystallography. L.F. and M.L. wrote the manuscript with input from all co-authors.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0330-9.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0330-9.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.L. or L.F.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Expression and purification. The coding sequence of *Metarhizium acridum* MCU without mitochondrial signal peptide (residues 62–484) was codon optimized and cloned into a modified pNGFP-BC plasmid with an N-terminal His-tag, GFPuv, and a 3C protease cleavage site. Based on limited proteolysis and secondary structure prediction, the unstructured N- and C- termini and a disordered loop at the NTD were removed. A construct lacking residues 62–98, 190–205, and 427–484 yielded weakly diffracting crystals. To promote better crystal growth, *Clostridium pasteurianum* rubredoxin was fused into a loop between S213 and G214 at the NTD, and two point mutations (H327W and M330F) were introduced. This construct, named MaMCU-Rub, was fully functional and produced improved crystals. To prepare MaMCU soluble domain, the N- terminal and C-terminal parts of the protein were directly fused together by deleting the transmembrane (TM) domain (Δ295–386). A disordered long loop in the NTD (residues 193–222) was further truncated. The final construct was named MaMCUsd and cloned into pNGFP-BC. Similarly, the coding sequence of *Fusarium graminearum* MCU was codon optimized and cloned into pNGFP-BC. The mitochondrial signal peptide and disordered N- and C termini were truncated (Δ1–120 and Δ448–480). Two disordered and non-conserved loops (residues 148–159 and 215–241) within the NTD were also removed.

To prepare the MaMCU protein sample, the transformed *E. coli* BL21 (DE3) cells were grown at 37 °C until the OD_{600} reached 0.8–1.0 and the temperature was lowered to 20 °C. Overexpression was induced by 0.2 mM isopropyl β-D-thiogalactoside and the cells were harvested 18 h post induction. Cells were lysed by sonication in a buffer containing 50 mM Tris, pH 8.0, 150 mM NaCl, 2 mM $CaCl_2$, and protease inhibitor cocktail. Membrane proteins were extracted by incubating the lysate or the isolated membrane with 2% n-dodecyl-β-D-maltoside (DDM, Anatrace) at 4 °C for 2 h. After removal of the insoluble fractions, the supernatant was incubated with prewashed cobalt resin at 4 °C for 2 h. After washing the resin with five column volumes (CV) of buffer A containing 20 mM Tris, pH 8.0, 150 mM NaCl, 2 mM $CaCl_2$, and 0.05% DDM supplemented with 10 mM imidazole and 10 CV of buffer A supplemented with 20 mM imidazole, the resin was washed with buffer A and then incubated with 3C protease overnight to cleave the tag. MaMCU was further purified by gel filtration (Superdex 200 increase, GE healthcare) in buffer containing 20 mM Tris, pH 8.0, 150 mM NaCl, 2 mM $CaCl_2$, and 0.03% DDM. The peak fraction was collected and concentrated to 10–15 mg/ml for further use.

MaMCU-Rub was purified similarly with a few modifications. The cell pellets were suspended in lysis buffer containing 50 mM Tris, pH 8.0, 150 mM NaCl, 10 mM $MgCl_2$, and protease inhibitor cocktail. After sonication, membrane proteins were extracted by 2% n-decyl-β-D-maltoside (Anatrace) at 4 °C for 2 h. The insoluble fractions were removed by centrifugation and the supernatant was incubated with prewashed cobalt resin at 4 °C for 2 h. The protein was purified by affinity chromatography using cobalt resin, released from the resin by 3C protease, and further purified by gel filtration (Superdex 200 increase, GE Healthcare) in buffer containing 10 mM HEPES, pH 7.5, 150 mM NaCl, 10 mM $MgCl_2$, and 0.1% n-undecyl-β-D-maltoside (Anatrace). The peak fraction was collected and concentrated to 10 mg/ml for crystallization. Expression of selenomethionine-labelled MaMCU-Rub protein was achieved by following a published protocol[47]. The protein was purified in the same manner as the unlabelled protein. To produce mercury-labelled MaMCU-Rub protein, cysteine point mutations were introduced by two-step PCR[48], and the protein was incubated with 2.5 mM thimerosal at room temperature for 2 h before further purification by gel filtration.

The soluble matrix domain of MaMCU, MaMCUsd, was expressed and purified following the same protocol as described above, except that surface reductive methylation was performed after elution from the affinity column[49] and the gel filtration buffer did not contain any detergent. The peak fraction was collected and concentrated to 10 mg/ml for crystallization trials.

FgMCU proteins were purified following the same protocol above for MaMCU. The final gel filtration buffer contained 20 mM Tris, pH 8.0, 150 mM NaCl, 2 mM $CaCl_2$, and 0.03% DDM.

Nanobody selection and purification. Nanobodies were obtained from screening immunized and synthetic libraries following published protocols[50,51]. Phage display libraries that contain about 10^8 transformants were used to screen specific binders to MCUs. After four rounds of selection, specific binders against MCUs were clearly enriched. The enriched clones were transferred into a yeast display system and further screened against MaMCU for three additional rounds following published protocols[52]. Nanobodies that specifically recognize MaMCU were cloned into expression vector pMES4 with His tag. Five nanobodies were expressed in *E. coli* BL21 (DE3) and purified according to the standard procedure using cobalt resin.

Crystallization. Crystallization trials were performed using the hanging-drop vapour-diffusion method at 20 °C. MaMCU-Rub protein was mixed at a 1:0.8 ratio with reservoir solution containing 100 mM HEPES, 7.5, 0.1 M $MgCl_2$, and 14% PEG 2000 MME. The crystals grew to full size in one to two weeks. Crystals were incubated in a cryoprotectant solution with the addition of 20% glycerol, and were flash-frozen in liquid nitrogen. MaMCU in complex with nanobody was prepared by mixing both components at a 1:1 ratio. The crystals were grown by vapour-diffusion in the presence of 15% PEG 2050, 200 mM $(NH_4)_2SO_4$, and 100 mM Na-cacodylate, pH 6.0. The crystals grew to full size in about one week. The crystals were cryo-protected by the addition of 20% glycerol and flash frozen in liquid nitrogen. The MaMCUsd crystals were grown using the same method as described above. Plate-like crystals were grown in 0.1 M glycine, pH 9.3, 15.5% PEG 4000, and 0.6 M $NaNO_3$. The crystals were cryo-protected by adding 24% ethylene glycol into the crystallization buffer, and were then flash-frozen in liquid nitrogen.

Data collection and structural determination. The diffraction data were collected at beamlines 24ID-C/E, 23ID-B/D and 17ID of the Advanced Photon Source, 5.0.2 of the Advanced Light Source, and 12-2 of the Stanford Synchrotron Radiation Lightsource. Native data set for MaMCU-Rub, and MaMCUsd was collected based on a single crystal and native data set for MaMCU was merged from three crystals. The diffraction data were processed using XDS[53] or HKL2000[54]. The selenomethionine anomalous data were merged from multiple crystals. The data were integrated using XDS[53] as implemented in RAPD (https://github.com/RAPD/RAPD). The mercury anomalous data were collected on multiple crystals or multiple spots on the same crystals and were processed by XDS[53] or HKL2000[54]. The initial phase for MaMCUsd was obtained using a 10 Å EM map as a search model using Phaser[55], and the phase was gradually extended through NCS averaging to a 3.1 Å resolution by RESOLVE through Phenix[56]. The map was of sufficient quality to trace secondary structures. The initial model was built using AutoBuild in Phenix. The model was manually rebuilt in Coot[57] and refined by Phenix[56] in an iterative manner. The initial single wavelength anomalous dispersion (SAD) phasing for MaMCU-Rub (crystal form I) was obtained from either selenomethionine anomalous data or mercury anomalous data using SHELX C/D/E in HKL2Map[58]. In addition, an MR-SAD (combination of molecular replacement and single wavelength anomalous dispersion) protocol was carried out for the selenomethionine anomalous data and mercury anomalous data based on a molecular replacement (MR) solution with the soluble domain model (MaMCUsd) as the search model. SAD and MR-SAD maps allowed tracing and registering the soluble domain with the aid of the MaMCUsd model. For the TM domain, initially ideal poly-alanine helices were manually placed into the SAD and MR-SAD maps, guided by the EM map. To assist assignment of the helix register, we made single cysteine mutants for each residue on the TM helices. The accessibility of cysteine for the mercury labelling was assessed by 5,5′-Dithiobis(2-nitrobenzoic acid) on the purified protein. For each mutant, the MR solution was obtained by Phaser[55]. The Hg sites were identified by Phaser EP in Phenix[56] and confirmed by anomalous difference maps. The register of the TM helices were aided by 12 Hg markers on two TM helices, one Hg marker on the linker connecting TM1 and TM2, as well as a selenomethione marker on TM2. For MaMCU in complex with the nanobody (crystal form II), the combined and scaled data from 3 crystals was anisotropically truncated using UCLA Diffraction Anisotropy Server. The data (3.9 Å × 3.1 Å × 3.3 Å along a*, b*, and c* axes, respectively) was used for refinement. The molecular replacement solution was found using MaMCU-Rub as a search model by Phaser[55]. The model was iteratively refined and rebuilt in Phenix[56], Buster[59] and Coot[57]. This final model was used for all structural interpretation on MaMCU. Structural quality was reported by MolProbity[60] and structural renderings were prepared using PyMOL (Schrödinger, LLC). Ion-conducting pore radius was calculated by HOLE[61].

Cryo-EM sample preparation. To reconstitute the MCU in amphipol, purified proteins after gel filtration were mixed with amphipol A8–35 (Anatrace) or PMAL-C8 (Anatrace) at a mass ratio of 1:3 for 3 h on ice. To prepare the MCU in nanodiscs, purified proteins were mixed with an equal molar ratio of the scaffold protein MSP1D1 in the presence of an 8 mM lipid mixture that contained 3:1 2-oleoyl-1-palmitoyl-sn-glycero-3-phosphocholine (POPC, Avanti) and cardiolipin (Avanti) for 2 h on ice. The protein samples with amphipol and nanodiscs were incubated with Bio-beads at 4 °C overnight to remove detergents and were further purified by gel filtration in a buffer containing 150 mM NaCl, 10 mM Tris, pH 8.0, and 2 mM $CaCl_2$. The peak fraction was collected for EM analysis.

Cell-based uptake assay. The MCU and a membrane-tethering calcium sensor Lck-GCaMP2[30,62] were co-expressed using a pETDuet1 vector. The transformed *E. coli* strain BL21(DE3) was induced as described above. A 3 ml cell culture was centrifuged and washed twice with 3 ml UAB buffer (100 mM KCl, 50 mM Tris, pH 7.5, 1 mM $MgCl_2$), and then diluted to an OD600 of 1.0 for an uptake assay. A large negative membrane potential across the *E. coli* plasma membrane[63] drives Ca^{2+} uptake. The assays were carried out using Greiner black polystyrene flat-bottom 96-well plates (Sigma-Aldrich). For each assay reaction, 50 μM or 0.5 mM calcium was added into 100 μl cells and immediately mixed to initiate the uptake. For ruthenium red inhibition experiments, 0.5–2 μM ruthenium red (Sigma-Aldrich) was mixed with cells before uptake assays. All experiments were performed at room

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

temperature and repeated at least four times. Data were recorded on an infinite M1000Pro plate reader and processed using Tecan i-contro software. Fluorescent signals were measured every 5–10 s with the excitation set at 485 nm and emission at 510 nm. Ca^{2+} uptake activity is quantified as a linear fit of the fluorescence signal during the first 20 s after the addition of calcium. To evaluate the expression level of the MCU mutants, MCU mutant constructs were cloned into pNGFP-BC. Proteins were expressed in 3 ml LB culture under the identical condition as for the functional assay. Solubilized proteins were analysed by SDS–PAGE and MCU-GFP fusion protein was visualized by in gel fluorescence using a Bio-Rad imaging system with excitation at 302 nm and emission at 440 nm.

Liposome-based flux assay. Liposome was reconstituted according to a published protocol[64]. Purified MCU proteins were mixed with *E. coli* polar lipids (Avanti) at a 1:80 ratio in the presence of reconstitution buffer (150 mM NaCl, 10 mM HEPES, pH 7.5, 5 mM EDTA) plus 0.12% (w/v) Triton-X100. Detergents were removed by overnight incubation with Bio-beads. After extrusion through a 0.4 μm membrane, proteoliposome was pelleted by ultracentrifugation at 100,000 g for 1 h. The proteoliposome was washed twice with ice-cold FAB buffer that contained 150 mM NMDG, 10 mM HEPES, pH 7.5, 5 mM EDTA and 2 μM ACMA, and then resuspended in the same buffer to a final lipid concentration of 100 mg/ml. The flux assay was performed following published protocols[31,65,66] in a 96-well format. A total of 2 μl of proteoliposomes were diluted into 100 μl of FAB. After 150 s, 1 μM carbonyl cyanide m-chlorophenyl hydrazone (CCCP, Sigma-Aldrich) was added to initiate Na^+ flux. A total of 20 nM sodium ionophore monensin (Sigma-Aldrich) was added near the end of the experiments to release Na^+ from the liposomes. For Gd^{3+} blockage experiments, 0.3 mM $GdCl_3$ was added to the FAB buffer before flux assays. Data collected on an infinite M1000Pro plate reader were processed using Tecan i-contro software. Fluorescent signals were measured every 15 s (excitation: 410 nm and emission: 490 nm).

Cysteine crosslinking. Double cysteine substitutions were introduced in a cysteine-free version of MaMCU by two-step PCR[48]. The mutants were expressed in the same manner as the WT protein. For crosslinking in detergent solution, 1 mg/ml purified protein in DDM buffer (20 mM Tris, pH 8.0, 150 mM NaCl, 2 mM $CaCl_2$, 0.03% DDM) was incubated with 0.1 mM of Cu(II) (1,10-phenantroline)$_3$ at 4 °C. For crosslinking in membrane, the membrane fraction containing overexpressed His-tagged protein was prepared and incubated with 0.3 mM of Cu(II) (1,10-phenantroline)$_3$ at 4 °C. The reactions were stopped at different time points by the addition of EDTA to a final concentration of 50 mM, and the reaction products were analysed by SDS–PAGE. For western blot analysis, the proteins were transferred from a gel to a PVDF membrane and probed with mouse anti-penta-His antibody (Qiagen), followed by detection with donkey anti-mouse IR680 secondary antibody (LI-COR Biosciences). Blots were visualized and analysed using a LI-COR Odyssey system. Double cysteine substitutions of mosquito MCU was expressed and purified in DDM with 0.01% cholesteryl hemisuccinate similarly to MaMCU. The crosslinking was performed in DDM buffer with 0.01% cholesteryl hemisuccinate in a similar manner to MaMCU.

Electron microscopy data acquisition. Negatively stained specimens were prepared following an established protocol with minor modifications[67]. Specifically, 2.5 μl of purified MCU was applied to glow-discharged copper EM grids covered with a thin layer of continuous carbon film, and the grids were stained with 2.5% (w/v) uranyl formate. These grids were imaged on a Tecnai T12 electron microscope (FEI) operated at 120 kV at a nominal magnification of 67,000x using a 4k × 4k CCD camera (UltraScan 4000, Gatan), corresponding to a calibrated pixel size of 1.68 Å on the specimen level. For cryo-EM of FgMCU in PMAL-C8 and nanodiscs, 4 μl of a 5 mg/mL sample containing 0.3 mM fluorinated Fos-Choline-8 (Anatrace) was applied to a glow-discharged Quantifoil holey carbon grid (1.2/1.3, 400 mesh), and blotted for 3 s with ~92% humidity before plunge-freezing in liquid ethane using a Cryoplunge 3 System (CP3, Gatan). For cryo-EM of MaMCU in A8–35, a 2.5 μl sample at a concentration of 2 mg/ml applied to a grid was similarly blotted and plunge-frozen using a CP3. The cryo-EM data were recorded on a 300 kV Polara electron microscope (FEI) at Harvard Medical School, and a 200 kV Talos Arctica and a 300 kV Titan Krios electron microscope (FEI) at the University of Massachusetts Medical School. All cryo-EM movies were recorded with a K2 Summit direct electron detector (Gatan) in super-resolution counting mode using UCSFImage4[68] or SerialEM[69]. The details of the EM data collection parameters are listed in Extended Data Table 1.

EM image processing. The EM data were processed as previously described[70]. Negative-stain EM images were binned over 2 × 2 pixels for further processing, yielding a pixel size of 3.36 Å. Dose-fractionated super-resolution movies collected using the K2 Summit direct electron detector were binned over 2 × 2 pixels, and then subjected to motion correction using the program MotionCor2[71]. A sum of all frames of each image stack was calculated following a dose-weighting scheme, and used for all image-processing steps except for defocus determination. The program CTFFIND4[72] was used to calculate defocus values of the summed images from all movie frames without dose weighting. Particle picking was performed using a semi-automated procedure with SAMUEL and SamViewer[73]. 2D classification of selected particle images was carried out by SPIDER (samclasscas.py) or by RELION[74,75]. Initial 3D models were generated with 2D class averages by SPIDER 3D projection matching refinement (samrefine.py), starting from a cylindrical density mimicking the size and shape of the MCU. 3D classification and refinement were carried out using "relion_refine_mpi" in RELION. The masked 3D classification focusing on the TM domain of FgMCU in nanodiscs with residual signal subtraction was performed following a previously described procedure[76]. The first round of 3D classification of all cryo-EM data sets was carried out without symmetry, which was followed by further 3D classification and refinement with two-fold symmetry. All refinements followed the gold-standard procedure, in which two half data sets were refined independently. The overall resolutions were estimated based on the gold-standard criterion of Fourier shell correlation (FSC) = 0.143. Local resolution variations were estimated from two half data maps using ResMap[77]. The amplitude information of the final maps were corrected by applying a negative B-factor using the program bfactor.exe[78]. To compare the crystal structure with the cryo-EM maps, the atomic model was fitted into the cryo-EM maps as a rigid body using the function of "Fit in Map" in UCSF Chimera[79].

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

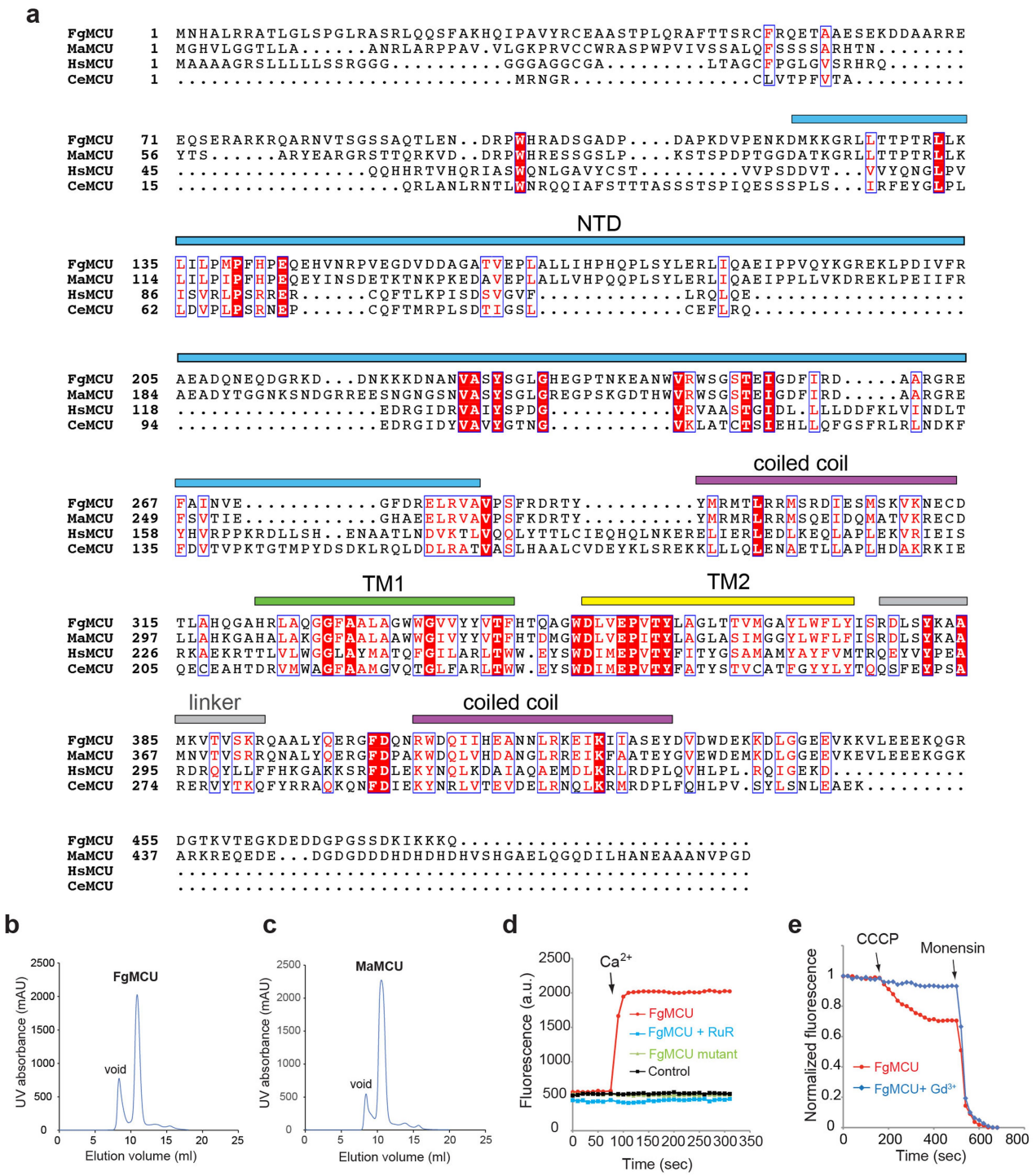
Data availability. The crystallographic coordinates and structural factors are deposited into the Protein Data Bank with accession numbers 6C5W and 6C5R. The 3D cryo-EM density maps of MaMCU in A8–35, FgMCU in PMAL-C8 and FgMCU in nanodiscs (overall, type 1 and type 2 conformations) have been deposited into the Electron Microscopy Data Bank under accession numbers EMDB-7800, EMDB-7801, EMDB-7802, EMDB-7803 and EMDB-7804.

47. Doublie, S. Preparation of selenomethionyl proteins for phase determination. *Methods Enzymol.* **276**, 523–530 (1997).
48. Ho, S. N., Hunt, H. D., Horton, R. M., Pullen, J. K. & Pease, L. R. Site-directed mutagenesis by overlap extension using the polymerase chain reaction. *Gene* **77**, 51–59 (1989).
49. Walter, T. S. et al. Lysine methylation as a routine rescue strategy for protein crystallization. *Structure* **14**, 1617–1622 (2006).
50. Pardon, E. et al. A general protocol for the generation of nanobodies for structural biology. *Nat. Protocols* **9**, 674–693 (2014).
51. McMahon, C. et al. Yeast surface display platform for rapid discovery of conformationally selective nanobodies. *Nat. Struct. Mol. Biol.* **25**, 289–296 (2018).
52. Chao, G. et al. Isolating and engineering human antibodies using yeast surface display. *Nat. Protocols* **1**, 755–768 (2006).
53. Kabsch, W. XDS. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010).
54. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
55. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
56. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
57. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of *Coot*. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
58. Pape, T. & Schneider, T. R. *HKL2MAP:* a graphical user interface for macromolecular phasing with SHELX programs. *J. Appl. Crystallogr.* **37**, 843–844 (2004).
59. Bricogne, G. et al. BUSTER (Global Phasing Ltd, Cambridge, UK, 2017).
60. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 12–21 (2010).
61. Smart, O. S., Neduvelil, J. G., Wang, X., Wallace, B. A. & Sansom, M. S. HOLE: a program for the analysis of the pore dimensions of ion channel structural models. *J. Mol. Graph.* **14**, 354–360, 376 (1996).
62. Tallini, Y. N. et al. Imaging cellular signals in the heart in vivo: Cardiac expression of the high-signal Ca^{2+} indicator GCaMP2. *Proc. Natl Acad. Sci. USA* **103**, 4753–4758 (2006).
63. Felle, H., Porter, J. S., Slayman, C. L. & Kaback, H. R. Quantitative measurements of membrane potential in *Escherichia coli*. *Biochemistry* **19**, 3585–3590 (1980).
64. Veenhoff, L. M. & Poolman, B. Substrate recognition at the cytoplasmic and extracellular binding site of the lactose transport protein of *Streptococcus thermophilus*. *J. Biol. Chem.* **274**, 33244–33250 (1999).
65. Zhang, J., Feng, Y. & Forgac, M. Proton conduction and bafilomycin binding by the V_0 domain of the coated vesicle V-ATPase. *J. Biol. Chem.* **269**, 23518–23523 (1994).
66. Feng, L., Campbell, E. B., Hsiung, Y. & MacKinnon, R. Structure of a eukaryotic CLC transporter defines an intermediate state in the transport cycle. *Science* **330**, 635–641 (2010).
67. Booth, D. S., Avila-Sakar, A. & Cheng, Y. Visualizing proteins and macromolecular complexes by negative stain EM: from grid preparation to image acquisition. *J. Vis. Exp.* **58**, 3227 (2011).
68. Li, X., Zheng, S., Agard, D. A. & Cheng, Y. Asynchronous data acquisition and on-the-fly analysis of dose fractionated cryoEM images by UCSFImage. *J. Struct. Biol.* **192**, 174–178 (2015).
69. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

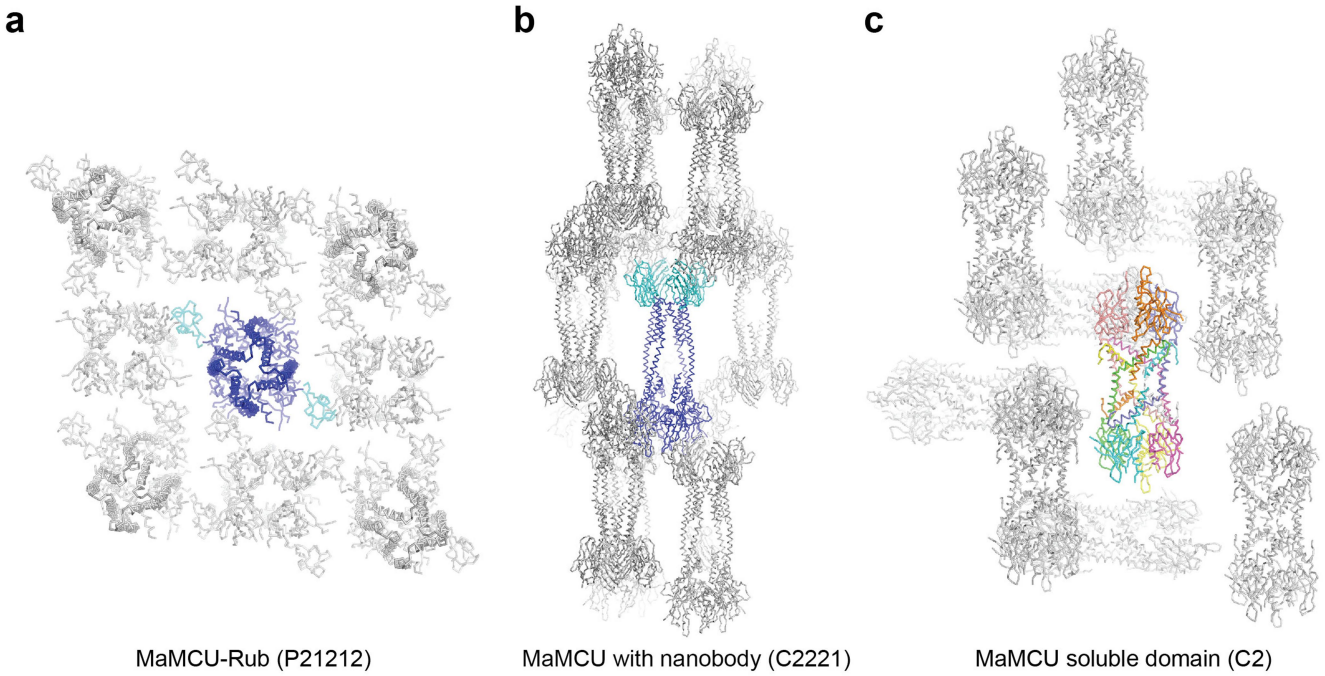
70. Mi, W. et al. Structural basis of MsbA-mediated lipopolysaccharide transport. *Nature* **549**, 233–237 (2017).
71. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
72. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).
73. Ru, H. et al. Molecular mechanism of V(D)J recombination from synaptic RAG1–RAG2 complex structures. *Cell* **163**, 1138–1152 (2015).
74. Scheres, S. H. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
75. Scheres, S. H. A Bayesian view on cryo-EM structure determination. *J. Mol. Biol.* **415**, 406–418 (2012).
76. Bai, X. C., Rajendra, E., Yang, G., Shi, Y. & Scheres, S. H. Sampling the conformational space of the catalytic subunit of human γ-secretase. *eLife* **4**, e11182 (2015).
77. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).
78. Lyumkis, D., Brilot, A. F., Theobald, D. L. & Grigorieff, N. Likelihood-based classification of cryo-EM images using FREALIGN. *J. Struct. Biol.* **183**, 377–388 (2013).
79. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).

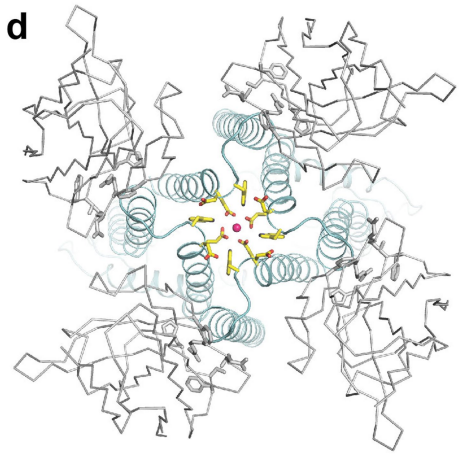
© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

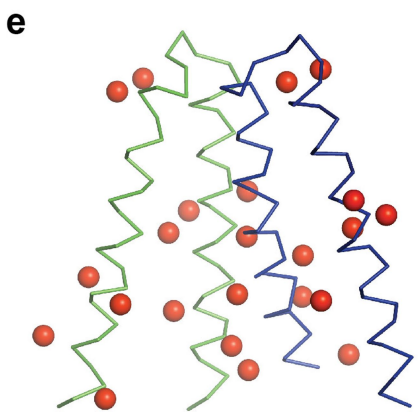


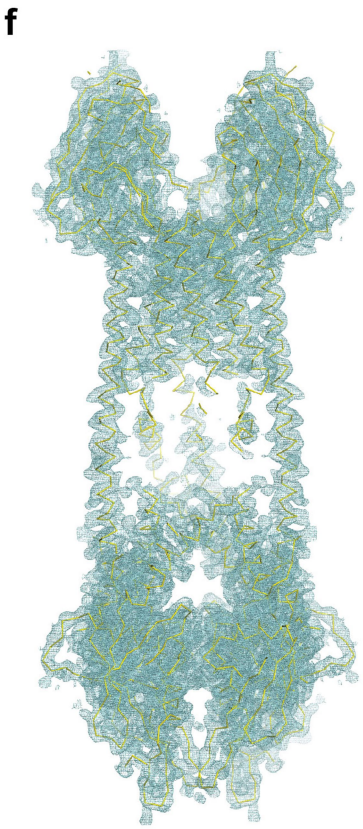
Extended Data Fig. 1 | Sequence alignment, sample preparations and functional characterizations. a, Sequence alignment of MCU. The protein sequences of MCUs from *Fusarium graminearum* (FgMCU), *Metarhizium acridum* (MaMCU), *Homo sapiens* (HsMCU) and *Caenorhabditis elegans* (CeMCU) were aligned using Clustal Omega with manual adjustment. The domain organization is shown above the sequences. **b**, FgMCU in DDM on size-exclusion chromatography. **c**, MaMCU in DDM on size-exclusion chromatography. Data in **b**, **c** are representative of five independent experiments with similar results. **d**, Ca^{2+} uptake activity of FgMCU in *E. coli*. Fluorescence changes upon addition of 0.5 mM Ca^{2+} are shown for wild-type (red), D351A/E354A (green), control (black), and with ruthenium red (blue). **e**, Cation flux activity of FgMCU in reconstituted liposomes. Fluorescence changes are shown for FgMCU (red) and with Gd^{3+} (blue). Data in **d**, **e** are representative of three independent experiments with similar results.

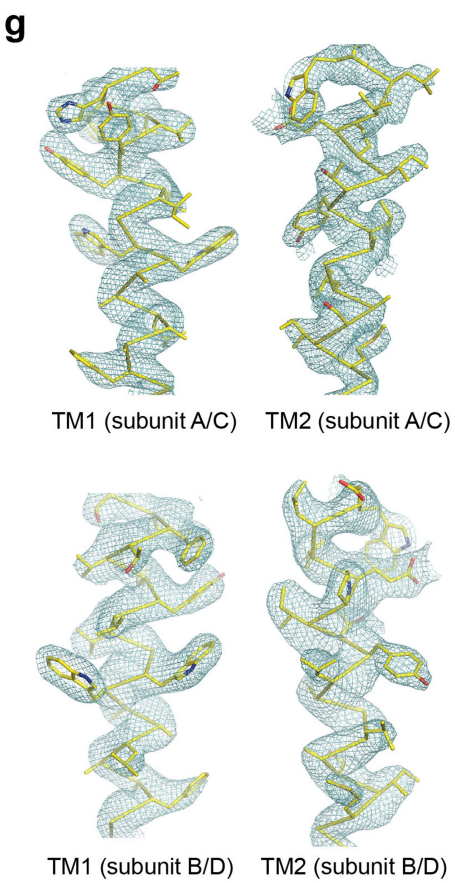
© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.









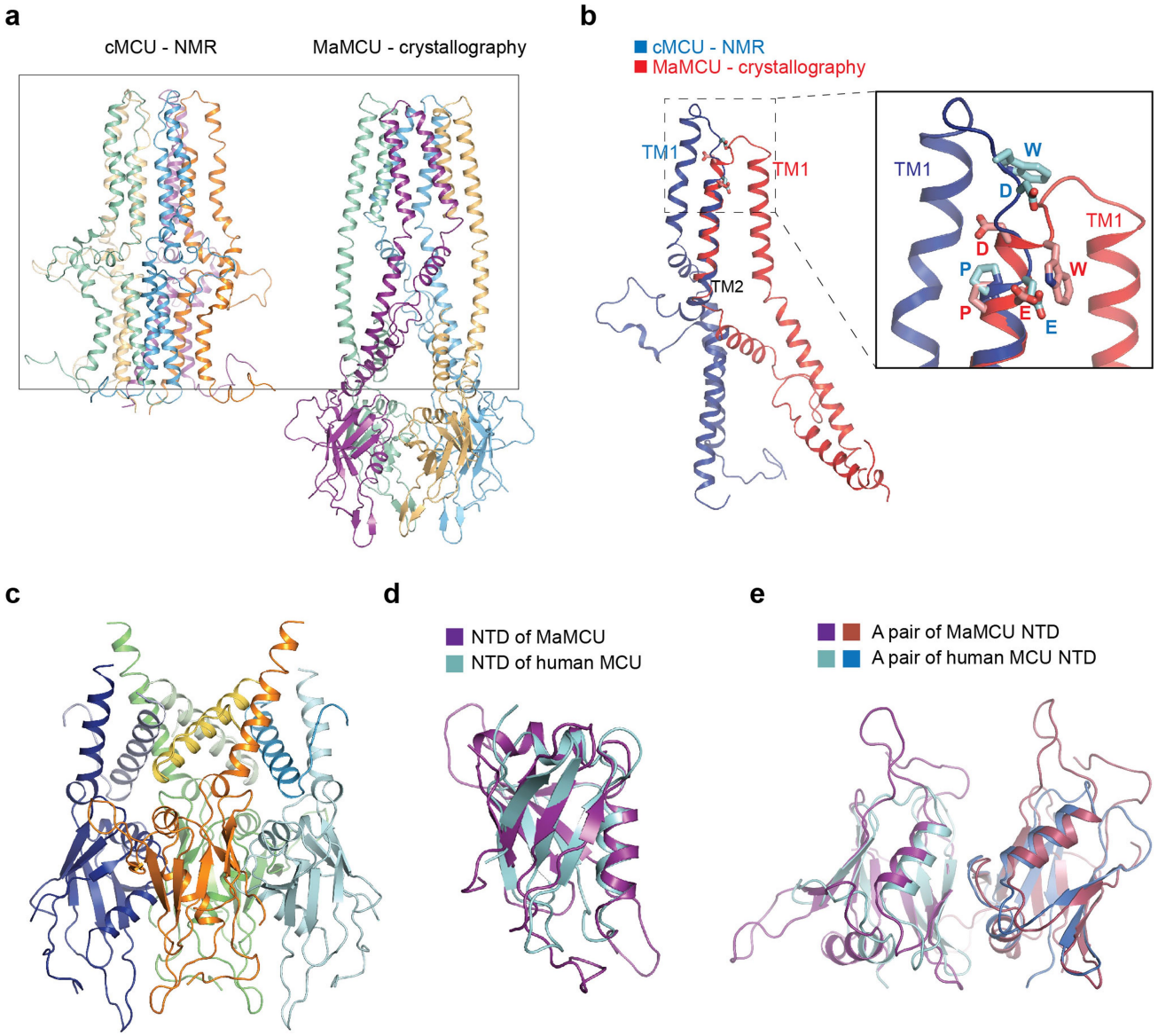


Extended Data Fig. 2 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

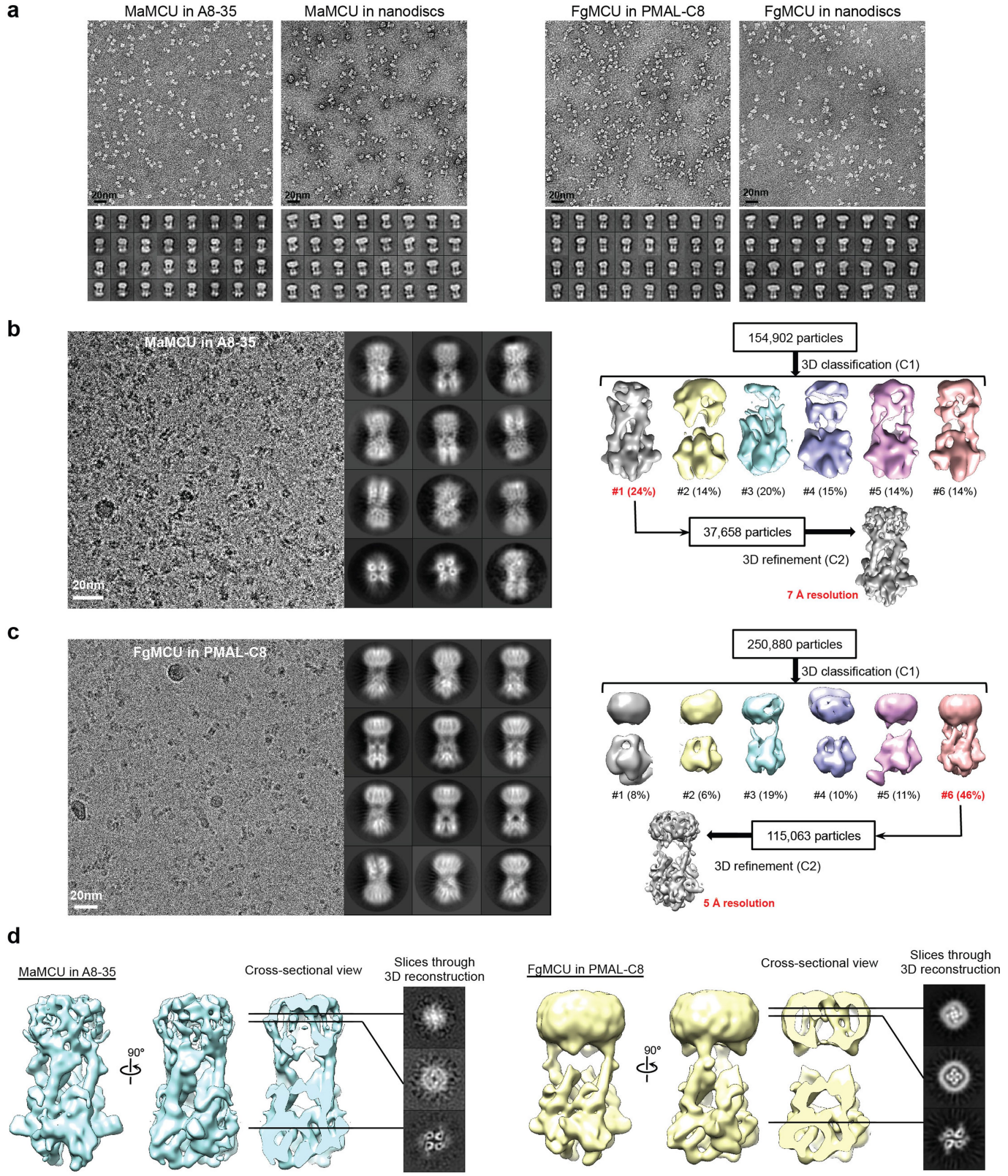
Extended Data Fig. 2 | Crystal packing and molecule interactions in the crystal. a, The crystal lattice structure of MaMCU-Rub. One MaMCU channel is coloured blue and its associated rubredoxin is coloured cyan. In the crystal, the transmembrane domain lacks clear protein-mediated contacts. Presumably, relatively disordered protein regions or detergent micelles mediate the contact. **b**, The crystal lattice structure of MaMCU in complex with nanobody. One MCU is shown in blue and the bound nanobodies are shown in cyan. **c**, The crystal lattice structure of the MaMCU soluble domain, shown in ribbons. Each chain is coloured differently. Two helices that formed a coiled-coil were derived from two different chains, in a domain-swapped manner. **d**, Close-up view of the interaction between nanobody and McMCU in the crystal. Nanobody binds near the end of TM1 and is distant from the selectivity filter (shown as coloured sticks). Nanobody is coloured grey and MaMCU is colored cyan. **e**, Mercury and selenium sites identified by MR-SAD. Two protomers of MaMCU in one asymmetric unit are shown as ribbons in green and blue, respectively. The mercury sites identified from 'marker mutants' are shown as red spheres; one selenium site from the labelled wild-type protein is shown in the same manner. **f**, $2F_o-F_c$ electron density map of MaMCU (cyan mesh, 1σ) with the structure model shown in ribbon. **g**, Close-up view of $2F_o-F_c$ electron density map of individual transmembrane helices (cyan mesh, 1.5σ) with side chains shown in sticks.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



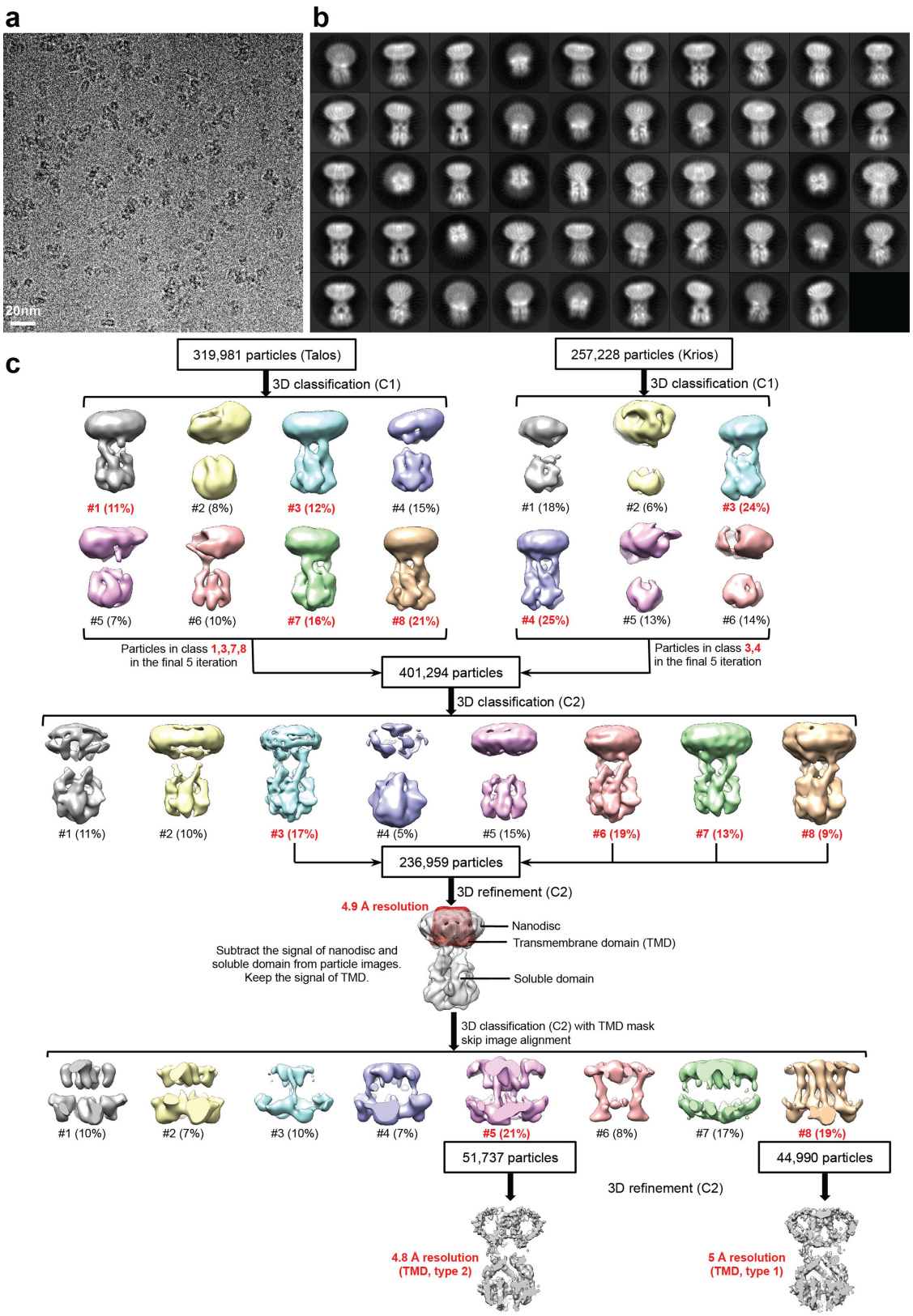
Extended Data Fig. 3 | Comparison of the NMR structure of cMCU with the crystal structure of MaMCU, and the structure of the soluble domain of MaMCU. a, Left, pentameric NMR structure of cMCU. Right, dimer-of-dimers structure of MaMCU. Box outlines the equivalent region including the transmembrane domain and coiled-coil. **b**, Comparison of a subunit of cMCU by NMR with that of MaMCU by crystallography. The helical part of the pore-lining TM2 is superimposed. The rest of the protein shows markedly different structure. Right, expanded view of the selectivity filter region. The most conserved WDxxEP residues (shown as sticks) are organized differently in these two structures. **c**, Ribbon representation of the soluble domain of MaMCU. Each subunit is coloured in blue, green, gold, or cyan, with the portions before and after the transmembrane domain of each subunit in slightly different colours. **d**, Structural comparison of the NTDs of MaMCU (purple) and human MCU (cyan). The two structures have the same fold and superimpose well onto each other. **e**, A pair of human MCU NTDs (cyan and blue) superimposes well onto one of dimers formed by MaMCU NTDs (purple and pink).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



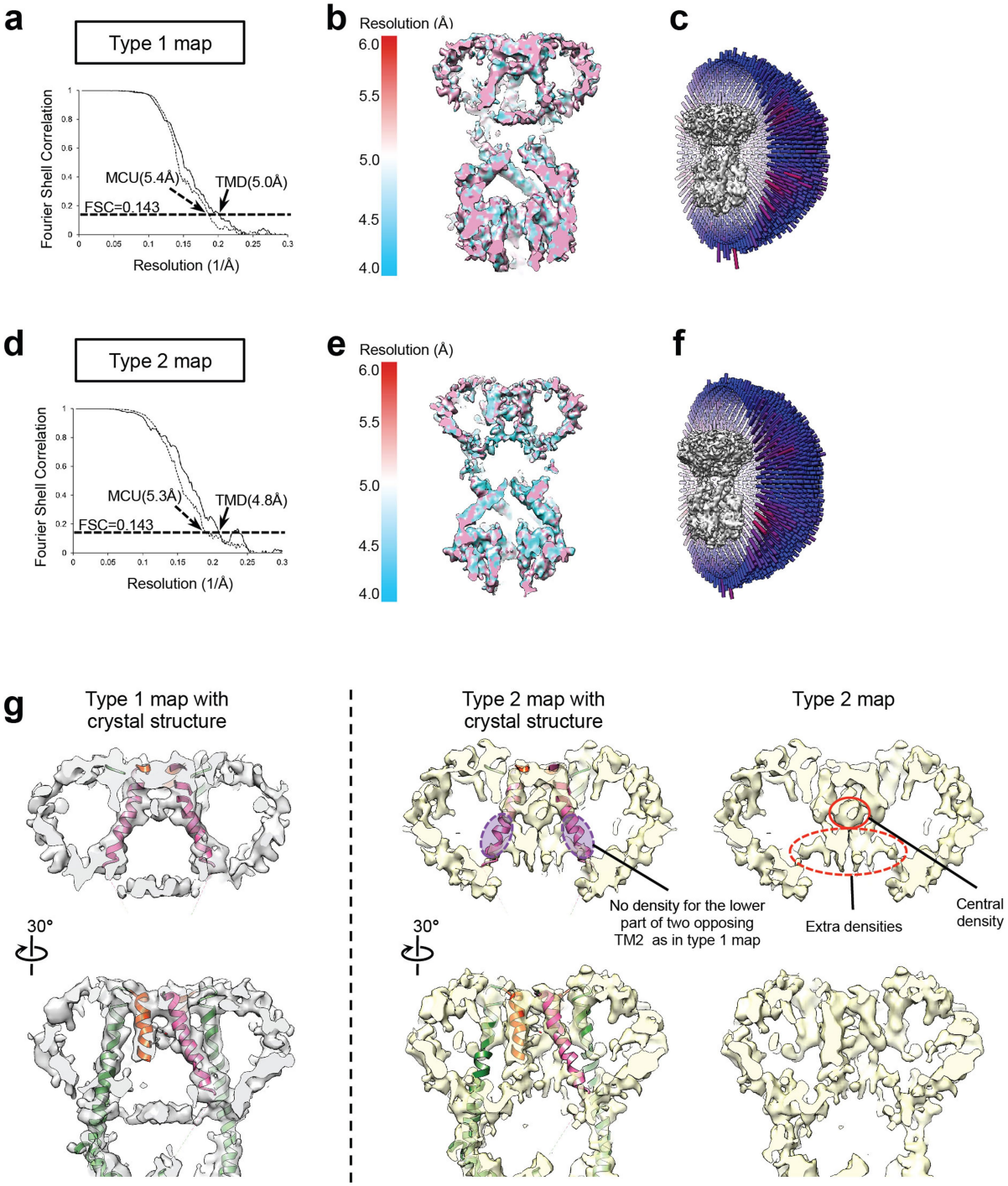
Extended Data Fig. 4 | Negative-stain EM and cryo-EM of MaMCU and FgMCU. a, Representative negative-stain EM image and 2D averages of MaMCU in A8–35 (from 28 micrographs with similar results), MaMCU in nanodiscs (from 28 micrographs with similar results), FgMCU in PMAL-C8 (from 20 micrographs with similar results), and FgMCU in nanodiscs (from 40 micrographs with similar results). The box dimension of the 2D averages is 215 Å. **b**, Representative cryo-EM image (from 1,862 micrographs with similar results), 2D averages, and image processing workflow of MaMCU in A8–35. The box dimension of the 2D averages is 172 Å. **c**, Representative cryo-EM image (from 4,055 micrographs with similar results), 2D averages, and image processing workflow of FgMCU in PMAL-C8. The box dimension of the 2D averages is 176 Å. **d**, Different views of the cryo-EM maps of MaMCU in A8–35 and FgMCU in PMAL-C8, both low-pass filtered to 7 Å. 2D averages and map calculation were not repeated.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



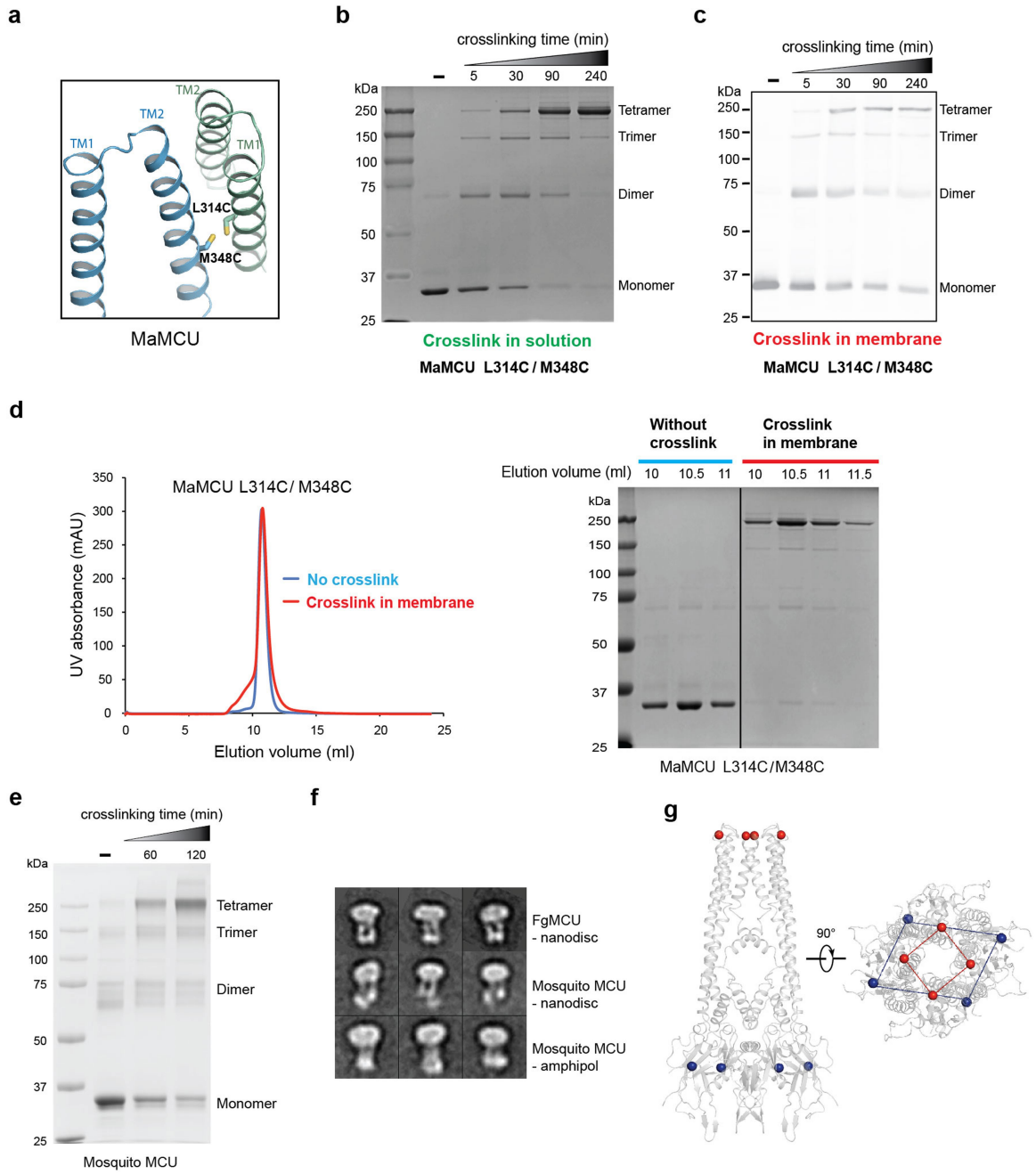
Extended Data Fig. 5 | Single-particle cryo-EM analysis of FgMCU in nanodiscs. a, Representative cryo-EM image (from 5,332 micrographs with similar results). **b**, 2D averages of cryo-EM particle images. The box dimension is 187 Å. **c**, Image processing workflow. Two cryo-EM data sets were first independently subjected to 3D classification without symmetry, and the particle images in good classes were combined and further processed with two-fold symmetry, generating a 4.9 Å-resolution cryo-EM map. Subsequently, a final round of 3D classification focused on the transmembrane domain sorted out two major classes, which generated two final maps at 5.0 and 4.8 Å resolution for the transmembrane domain. The conformations in these two maps are designated as type 1 and type 2. 2D averages and map calculations were not repeated.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



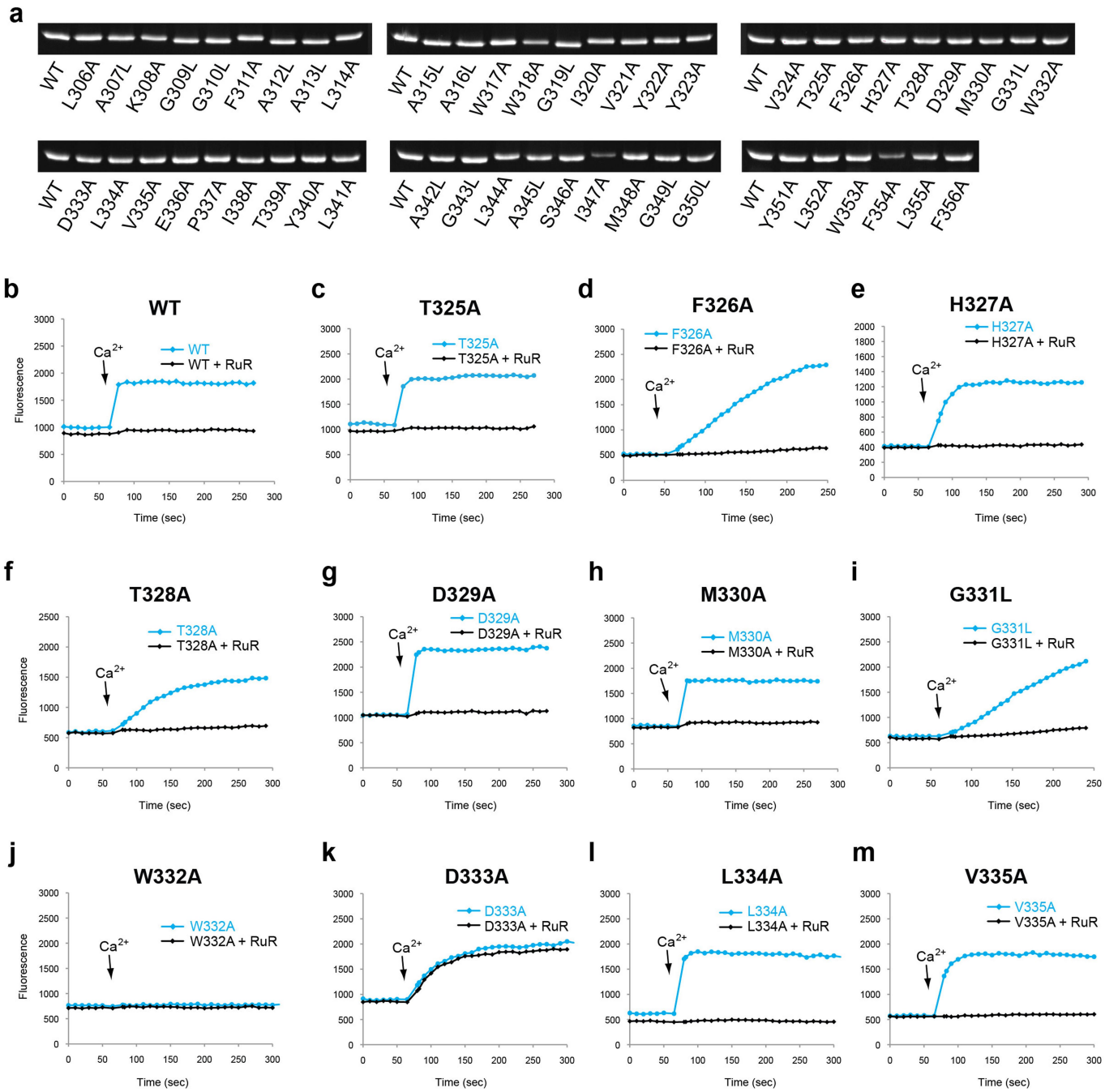
Extended Data Fig. 6 | Cryo-EM maps of two conformations of FgMCU in nanodiscs. a, Gold-standard FSC curves of the type 1 final cryo-EM map, calculated with a soft mask to include only the transmembrane domain (TMD, solid line) or the entire MCU molecule without nanodisc (dashed line). The estimated resolutions based on an FSC = 0.143 criterion are indicated. **b**, Cross-sectional view of the type 1 final cryo-EM map filtered to 5.0 Å resolution and coloured according to local resolution. **c**, Angular distribution of the particle images used for calculating the type 1 final cryo-EM map. **d**–**f**, Same as **a**–**c**, but for the type 2 conformation. **g**, Two cross-sectional views of the transmembrane domains of the type 1 and type 2 cryo-EM maps, superimposed onto an MaMCU-based homology model of FgMCU. The missing density and extra densities in the type 2 map are indicated. Map calculations in **a** and **d** were not repeated.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 7 | Tetrameric assembly of MaMCU, overall architecture of a metazoan MCU and symmetry mismatch of MaMCU. **a**, The design of cysteine pair mutations for crosslinking transmembrane subunits. L314 on TM1 and M348 on TM2 were mutated to cysteines (side chains shown as sticks). **b**, Site-directed disulfide cross-linking of MaMCU in detergent solution. L314 on TM1 and M348 on TM2 were mutated to create a cysteine pair at the protomer interface on an otherwise cysteine-free background. The double cysteine mutant was purified and oxidized by copper phenanthroline (CuP) for various time durations. The cross-link product was analysed by SDS–PAGE. **c**, Site-directed disulfide cross-linking of MaMCU in the membrane. Isolated membrane from the MaMCU L314C/M348C mutant was treated with CuP for an increasing amount of time. The cross-linking product was analysed by western blot using an antibody against the His tag on the protein. **d**, The size-exclusion profile of MaMCU with or without crosslinking. MaMCU L314C/M348C mutant treated or untreated with CuP in the membrane was purified and analysed by size-exclusion chromatography. The peak fractions were subjected to SDS–PAGE and visualized by Coomassie blue staining. **e**, Site-directed disulfide cross-linking of mosquito MCU in detergent solution. A pair of residues (T294 on TM1 and V341 on TM2) that were equivalent to those used in MaMCU crosslinking at the protomer interface were mutated to cysteine. The purified double cysteine mutant was oxidized by CuP and analysed by SDS–PAGE. The tetramer position on the SDS–PAGE gel is the same as that of MaMCU in **b**. **f**, 2D averages of negatively stained FgMCU in nanodiscs, mosquito MCU (T294C/V341C) in nanodiscs, and mosquito MCU (T294C/V341C) in amphipol A8–35. 2D averages were not repeated. **g**, Geometry of markers in transmembrane domain and NTD of MaMCU. Site-specific mercury marker on transmembrane domain (red spheres) and Cα position of Ile115 in NTD (blue spheres) are indicated. Data in **b**–**e** are representative of three independent experiments with similar results. For gel source data, see Supplementary Fig. 1.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 8 | Expression levels of MaMCU mutants and ruthenium red sensitivity of selected mutants. a, The expression level of wild-type MaMCU and those with a point mutation in the transmembrane domain. Green fluorescent protein-tagged protein was analysed on SDS–PAGE and the protein was visualized by in-gel fluorescence. The experiments were repeated twice independently with similar results. For gel source data, see Supplementary Fig. 1. **b**–**m**, The effect of substituting residues near the intermembrane surface of MaMCU on its ruthenium red sensitivity. Ca^{2+} uptake activity of the wild-type or mutant MaMCU was assayed in an *E. coli*-based assay (0.5 mM Ca^{2+}, 1 μM RuR). The Ca^{2+} uptake in the absence (blue) or in the presence (black) of ruthenium red is reflected by fluorescence changes. The experiments were repeated three times independently with similar results.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 1 | Statistics of X-ray data collection and refinement

	MaMCU	MaMCU soluble domain	MaMCU-Rub	MaMCU Ca^{2+}
Data collection				
Space group	*C*2221	*C*2	*P*21212	*C*2
Cell dimensions				
a, *b*, *c* (Å)	88.0, 290.4, 126.1	261.8, 119.9, 88.0	85.8, 109.9, 170.1	140.5, 263.81, 124.9
α, β, γ (°)	90.00, 90.00, 90.00	90.00, 106.88, 90.00	90.00, 90.00, 90.00	90.00, 124.08, 90.00
Resolution (Å)*	43.88 -3.05 (3.16-3.05)#	50.00-3.10 (3.17-3.10)	50.00-3.23 (3.41-3.23)	106.7 -3.6 (3.80-3.60)
Wavelength (Å)	1.00000	1.00000	1.00000	1.4938
R_{merge} (%)*	26.5 (>100)	8.9 (>100)	32.8 (>100)	19.3 (>100)
$I / \sigma I$*	11.1 (1.2)	14.1 (1.0)	4.7 (1.2)	14.7 (1.9)
Completeness (%)*	99.9 (99.9)	100.0 (100.0)	99.7 (99.8)	98.6 (97.7)
Redundancy*	13.1 (13.0)	3.4 (3.6)	12.4 (13.8)	20.5 (18.7)
Refinement				
Resolution (Å)	45.0 - 3.10	41.75 - 3.10		
Unique reflections	23,520	47,294		
R_{work} / R_{free} (%)	25.5/30.1	26.1/30.6		
No. atoms				
Protein	5746	9966		
Ligand/ion	1	N.A.		
B-factors				
Protein	75.71	99.07		
Ligand/ion	32.96	N.A.		
R.m.s deviations				
Bond lengths (Å)	0.002	0.003		
Bond angles (°)	0.54	0.70		

*High-resolution shell is shown in parentheses.
#Reflections after anisotropic correction (3.9, 3.1, and 3.3 Å, respectively along a*, b*, and c* reciprocal cell directions) were used in the refinement.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 2 | Cryo-EM data collection statistics

	FgMCU nanodisc-overall (EMDB-7800)	FgMCU Nanodisc – type 1 (EMDB-7801)	FgMCU Nanodisc – type 2 (EMDB-7802)	FgMCU amphipol PMAL-C8 (EMDB-7803)	MaMCU amphipol A8-35 (EMDB-7804)
Data collection and processing					
Microscope	Talos/Krios	Talos/Krios	Talos/Krios	Polara	Polara
Magnification	36,000/130,000	36,000/130,000	36,000/130,000	39,000	31000
Voltage (kV)	200/300	200/300	200/300	300	300
Electron exposure (e–/Å^2)	72/46	72/46	72/46	74	46
Defocus range (μm)	0.5-2.5	0.5-2.5	0.5-2.5	1.5-3.0	1.0-3.0
Pixel size (Å)	1.169/1.09	1.169/1.09	1.169/1.09	0.98	1.23
Symmetry imposed	C2	C2	C2	C2	C2
Number of collected movies	2,459/2,873	2,459/2,873	2,459/2,873	4,055	1,862
Initial particle images (no.)	577,209	577,209	577,209	250,880	154,902
Final particle images (no.)	236,959	44,990	51,737	115,063	37,658
Map resolution (Å) (TMD)	4.9	5.4 (5.0)	5.3(4.8)	5.0	7.0
FSC threshold	0.143	0.143	0.143	0.143	0.143

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

ARTICLE

https://doi.org/10.1038/s41586-018-0331-8

Cryo–EM structures of fungal and metazoan mitochondrial calcium uniporters

Rozbeh Baradaran[1,3], Chongyuan Wang[1,3], Andrew Francis Siliciano[1,2] & Stephen Barstow Long[1]*

The mitochondrial calcium uniporter (MCU) is a highly selective calcium channel and a major route of calcium entry into mitochondria. How the channel catalyses ion permeation and achieves ion selectivity are not well understood, partly because MCU is thought to have a distinct architecture in comparison to other cellular channels. Here we report cryo–electron microscopy reconstructions of MCU channels from zebrafish and *Cyphellophora europaea* at 8.5 Å and 3.2 Å resolutions, respectively. In contrast to a previous report of pentameric stoichiometry for MCU, both channels are tetramers. The atomic model of *C. europaea* MCU shows that a conserved WDXXEP signature sequence forms the selectivity filter, in which calcium ions are arranged in single file. Coiled–coil legs connect the pore to N–terminal domains in the mitochondrial matrix. In *C. europaea* MCU, the N–terminal domains assemble as a dimer of dimers; in zebrafish MCU, they form an asymmetric crescent. The structures define principles that underlie ion permeation and calcium selectivity in this unusual channel.

The Ca^{2+} channel of the inner mitochondrial membrane, known as the mitochondrial calcium uniporter (MCU), enables mitochondria to take up substantial amounts of Ca^{2+} from the cytosol during intracellular Ca^{2+} signalling events[1–5]. Ca^{2+} influx stimulates mitochondrial ATP synthesis, controls the mitochondrial permeability transition that leads to cell death, and acts in a buffering role to shape cytosolic calcium signals[6]. A negative electric potential (approximately −160 mV) of the mitochondrial matrix relative to the cytosol, which is established by the proton gradient across the inner membrane, is the driving force for Ca^{2+} influx[7]. MCU is one of the most Ca^{2+}-selective channels known[8] and must be highly discriminatory to enable the uptake of Ca^{2+} while excluding other cytosolic cations, of which potassium is more than 100,000 times more abundant ($[Ca^{2+}]_{cytosol} \approx 0.1–1$ μM and $[K^+]_{cytosol} \approx 150$ mM). The selectivity of MCU is of paramount importance because substantial cation leakage could collapse the electromotive force used for ATP synthesis. It is thought that selectivity is achieved, in part, by high-affinity binding of Ca^{2+} (dissociation constant (K_d) ≤ 2 nM) within a 'selectivity filter' region of the pore[8]. The selectivity filter might involve a conserved signature sequence (WDXXEP) near the channel's two predicted transmembrane helices[3,4], but its 3D conformation is not known, partly because MCU shares no apparent sequence similarity to other ion channels. Ca^{2+} ions can move rapidly through the pore of MCU (maximally at about 5×10^6 Ca^{2+} s^{-1})[8], and this raises the apparent paradox of how both ion selectivity and high conductance are achieved.

MCU channels are present in most clades of eukaryotic life, including protists, plantae, amoebozoa, fungi, and metazoa[9]. The MCU protein subunit oligomerizes to form the Ca^{2+}-conducting pore[3–5], but channel activity is regulated differently among clades. Metazoans require the single-pass transmembrane protein EMRE for Ca^{2+} uptake, whereas the genomes of fungi, amoeba, and plants do not have a recognizable EMRE orthologue[10]. Many MCU channels are further regulated by MICU1 and MICU2 proteins, which are present in the intermembrane space[11,12], and by the protein MCUR1[13–15]. The structures of MCU from zebrafish and the fungus *C. europaea* presented here define the architecture of the channel and give insight into the mechanisms of ion permeation and Ca^{2+} selectivity.

Structure determination and oligomeric state

C. europaea MCU (CeMCU) and zebrafish MCU have 44% and 91% sequence similarity to human MCU, respectively, and were selected for structural determination on the basis of good biochemical stability in detergent-containing solutions (see Methods, Extended Data Fig. 1). Zebrafish MCU catalyses mitochondrial Ca^{2+} uptake when co-expressed with EMRE (Extended Data Fig. 2a–c). Uptake is blocked by ruthenium red and is dependent upon the proton gradient. Thus, zebrafish MCU recapitulates known properties of the human channel[3,16].

An orthologue of EMRE has not been identified in the genome of *C. europaea*. MCU channels from *Dictyostelium discoideum* and *Arabidopsis thaliana*, two other species without an EMRE, can mediate mitochondrial Ca^{2+} uptake when heterologously expressed[16,17]. However, we did not observe uptake when CeMCU or several other fungal MCU channels were expressed in yeast or mammalian cells, and conjecture that at least a subset of fungal channels may be regulated differently from other MCU channels.

For cryo-electron microscopy (cryo-EM) sample preparation, full length CeMCU and zebrafish MCU were produced using eukaryotic expression systems and purified using gentle, non-denaturing, methods. In a previous NMR study of a truncated version of *Caenorhabditis elegans* MCU lacking an N-terminal domain (NTD), which resulted in a different structure from the one presented here[18], a prokaryotic system was used for expression and purification involved a relatively harsh detergent (foscholine-14), which is rarely used in structural studies of membrane proteins[19].

The small molecular weight of MCU, just approximately 110 kDa for the ordered region of the channel assembly, poses challenges for structure determination by cryo-EM. We overcame these by collecting micrographs at high magnification and high total electron dose (see Methods). De novo initial 3D models of CeMCU and zebrafish

[1]Structural Biology Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [2]Weill Cornell/Rockefeller/Sloan Kettering Tri-Institutional MD-PhD Program, New York, NY, USA. [3]These authors contributed equally: Rozbeh Baradaran, Chongyuan Wang. *e-mail: longs@mskcc.org

© 2018 Springer Nature Limited. All rights reserved.

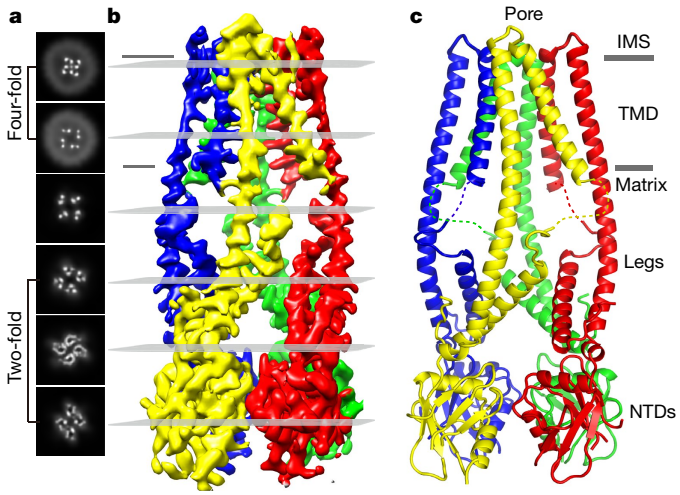


Fig. 1 | Cryo-EM map and overall structure of CeMCU. a, Slices of a 3D reconstruction generated without imposing symmetry (Extended Data Fig. 3). **b**, Final 3D reconstruction viewed parallel to the membrane (8σ contour level). Density for the detergent micelle is present at lower contours (for example, Extended Data Fig. 3). Subunits are coloured uniquely. **c**, Ribbon representation in the same orientation as **b**, with approximate membrane boundaries as grey bars. IMS, intermembrane space. Disordered regions connecting TM2 and α6 are indicated as dashed lines.

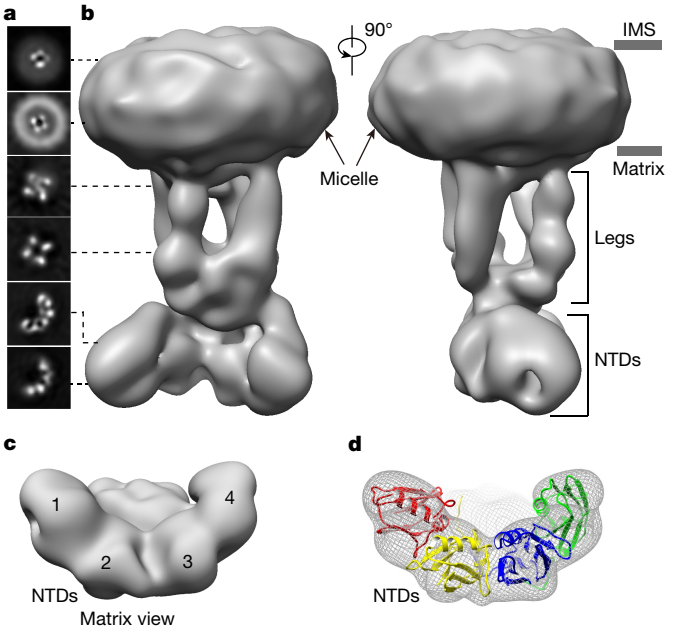


Fig. 2 | Cryo-EM reconstruction of zebrafish MCU. a, 2D slices of the 3D reconstruction of zebrafish MCU are shown at the indicated positions. **b**, Two orthogonal views of the 3D reconstruction viewed parallel to the membrane. Grey bars indicate approximate membrane boundaries. Density attributed to the detergent micelle surrounding the TMD is indicated. **c**, Matrix view, showing density for four NTDs. The TMD region is removed for clarity. **d**, Atomic models (PDB: 4XTB) of four NTDs are fit into the density from **c** (mesh).

MCU were generated independently and refined without imposing symmetry in order to avoid bias with regard to oligomeric state or overall structure (Extended Data Figs. 3, 4). The 3D reconstructions demonstrate that both channels assemble as tetramers of MCU subunits (Figs. 1, 2). 3D classification of the particle images, which would reveal global differences such as differences in oligomeric state, consistently yielded reconstructions that were tetrameric (Extended Data Figs. 3, 4). A tetrameric oligomeric state is supported by crosslinking experiments of purified protein and of human MCU within mitochondrial membranes (Extended Data Fig. 2h–j). The 3D reconstructions of CeMCU and zebrafish MCU were determined at overall resolutions of 3.2 Å and 8.5 Å, respectively (Figs. 1, 2 and Extended Data Figs. 5, 6). Density for most amino acid side chains of CeMCU allowed us to construct an atomic model with good stereochemistry and good correlation with the cryo-EM density (Fig. 1, Extended Data Fig. 6 and Extended Data Table 1). Because of the higher resolution, the discussion of atomic details that follows derives from CeMCU. Although individual α-helices cannot be resolved in the map of zebrafish MCU, aside from differently organized NTDs, the overall structures of zebrafish MCU and CeMCU are similar (Figs. 1, 2).

Overall architecture

MCU has a novel architecture for an ion channel (Fig. 1, Extended Data Fig. 7). The molecular surface contains a hydrophobic transmembrane domain (TMD) that would reside within the inner mitochondrial membrane in a cellular setting and a large hydrophilic region in the matrix (Extended Data Fig. 7d). The TMD has four-fold symmetry (Figs. 1, 2) with eight transmembrane helices (the two transmembrane helices (TM1 and TM2) from each of the subunits (Fig. 1)). The transmembrane helices interact through mostly hydrophobic amino acids (Extended Data Fig. 7g). Short 'turrets' connecting these helices are the only protrusions into the intermembrane space. On the matrix side, four 'legs' extend from the TMD. The legs are composed of an N-terminal extension of the TM1 helix ($TM1_{ext}$) and a C-terminal helix (α6) that forms a coiled-coil with it in the matrix (Fig. 1c, Extended Data Fig. 7a). Sequence analysis indicates that all MCU channels contain analogous coiled-coil regions[9]. The legs are separated from one another, creating ample space for ions to diffuse between them in the matrix. An NTD, which each MCU orthologue is predicted to contain[9], is located at the end of each of the legs, approximately 70 Å from the membrane. In CeMCU, the four NTDs are arranged as a dimer of dimers, and this yields an overall two-fold (*C*2) symmetry for the channel (Fig. 1). In zebrafish MCU, the four NTDs are arranged as an asymmetric crescent (Fig. 2). Thus, in both channels, there is a symmetry transition between the four-fold symmetric TMD and the matrix region. The cryo-EM structures differ substantially from the previous structural model of MCU that was deduced from NMR experiments[18] both on the subunit level and in terms of its assembly (Extended Data Fig. 8).

Ion pore

MCU contains a single ion pore in the TMD that is located along the symmetry axis and would be perpendicular to the membrane in a cellular setting (Figs. 1, 3). The secondary structure of the channel is entirely α-helical within the TMD; amino acid side chains emanating from α-helices form the walls of the pore. The pore can be divided into two portions: a narrower portion that occupies the outer leaflet of the membrane and contains the selectivity filter and a wider 'vestibule' portion located in the inner leaflet (Fig. 3). In the outer leaflet, TM2 lines the pore and TM1 is located at the periphery, whereas both TM1 and TM2 line the vestibule. The WDXXEP signature sequence (WDVMEP in CeMCU) adopts an α-helical secondary structure within TM2, forming approximately three turns of this helix with the W residue (Trp224) at its N-terminal end. The side chains of the D and E residues (Asp225 and Glu228) protrude into the pore. The W and P residues (Trp224 and Pro229) are tightly packed and generally oriented towards the pore, although they are not expected to interact with Ca^{2+}. The X residues (Val226 and Met227), face away from the pore and would be exposed to the membrane in a cellular setting (Fig. 4b). Pro229 induces a slight bend in the TM2 helix owing to local disruption of the α-helical hydrogen bonding (Fig. 4b).

To assess the relevance of the conformation of the WDXXEP sequence observed in CeMCU to human MCU, we introduced alanine and other mutations at each position of the signature sequence

© 2018 Springer Nature Limited. All rights reserved.

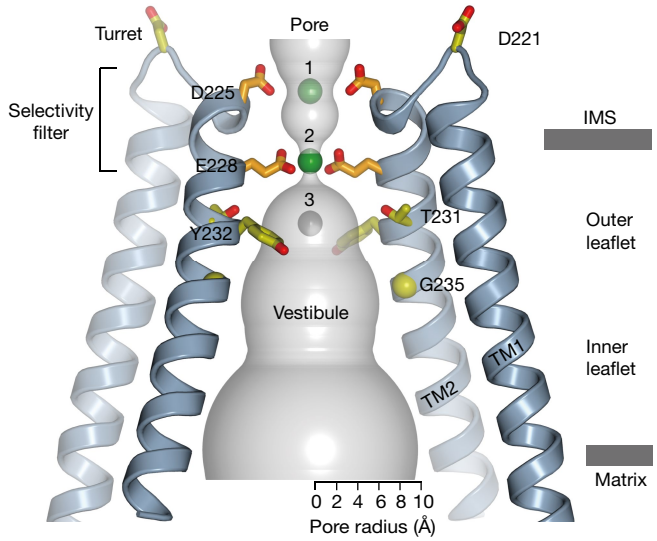


Fig. 3 | Ion pore. Within a ribbon representation of two subunits of CeMCU is a representation (grey surface) of the minimal radial distance from the centre of the pore to the nearest van der Waals protein contact. Selected amino acids are shown as sticks.

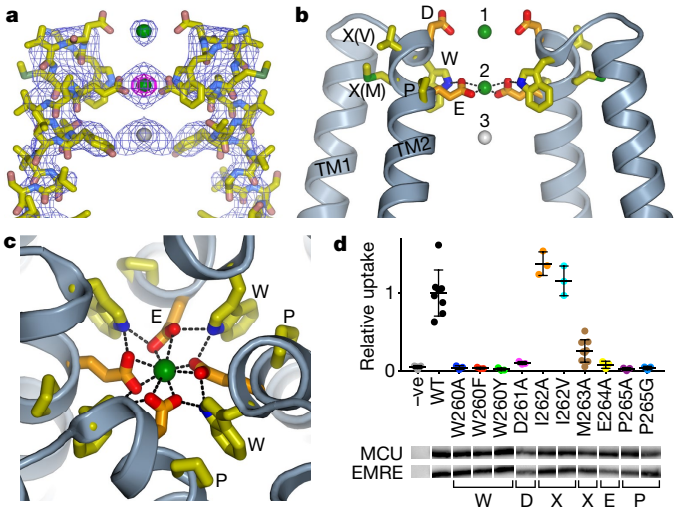


Fig. 4 | Selectivity filter. a, Density within the selectivity filter region of CeMCU. The map is contoured at 8σ (blue mesh) and 20σ (magenta mesh) and shown in the context of two TM2 helices (sticks; other regions are omitted for clarity). Green spheres indicate Ca^{2+} ions in sites 1 and 2; a grey sphere indicates site 3. **b**, Side view of the selectivity filter region, showing two subunits as ribbons and the WDXXEP signature sequence as sticks. Dashed lines indicate hydrogen bonds and Ca^{2+} coordination. **c**, Close up of site 2, oriented approximately 60° from the plane of the membrane, showing all four subunits. **d**, Mitochondrial Ca^{2+} uptake by indicated mutants of human MCU relative to wild type (WT) when co-expressed with EMRE (mean ± s.d., independent experiments: $n = 8$ for wild type, $n = 9$ for M263A, and $n = 3$ for the remainder, Extended Data Fig. 2d–g). Western blots demonstrate expression. −ve, untransfected cells.

of human MCU (WDIMEP) and evaluated mitochondrial Ca^{2+} uptake of the mutants by co-transfecting with wild-type human EMRE in mammalian cells in which the endogenous *MCU* and *EMRE* genes had been disrupted (MCU/EMRE knockout cells). Consistent with a pattern predicted by the structure, and with previous mutagenesis of the acidic amino acids[3], Ca^{2+} uptake was not observed when pore-oriented amino acids (W, D, E, or P) were mutated, but was observed for mutants of either X residue, which face the membrane in the structure of CeMCU (Fig. 4d and Extended Data Fig. 2d–g). The first X residue, Ile262 of human MCU, was particularly insensitive to mutation. The corresponding side chain of CeMCU (Val226) has no protein contacts in the structure. Both the I262V mutant, in which the signature sequence is converted to that of CeMCU, and the I262A mutant had full activity. M263A showed diminished but significant activity relative to the wild type, and this is consistent with the observation that while Met227 of CeMCU faces the membrane it also interacts with TM1 (Fig. 4b). These data support the conclusion that the signature sequence adopts similar conformations in metazoan and fungal MCU channels.

Selectivity filter

The selectivity filter is formed by symmetrical arrangement of WDXXEP signature sequences around the pore (Fig. 4). This region is both the narrowest portion of the pore and the only one that is lined by acidic amino acids. Its α-helical secondary structure is distinct from channels belonging to the superfamilies of CaV, NaV, Kv, TRP, RyR, and IP3R cation channels, which possess loop-like secondary structures within their selectivity filters[20–27]. The D (Asp225) residues of the tetrameric channel form an acidic mouth on the side of the pore facing the intermembrane space, giving the pore a radius of approximately 2.2 Å in this region (Fig. 3). The E (Glu228) residues form a second acidic ring, with a pore radius of less than 1 Å (Fig. 3). These glutamate residues extend fully away from the helical axis of TM2 and point towards the centre of the pore (Fig. 4c). The substitution of an aspartate residue would markedly change the pore dimensions. Accordingly, the corresponding E264D mutation of human MCU is defective in mitochondrial Ca^{2+} uptake[18]. Each W side chain (Trp224) is oriented down such that it interdigitates with the E (Glu228) side chains and appears to stabilize the conformation of the selectivity filter from behind (Fig. 4b, c). Each W side chain also packs against the P (Pro229) residue from a neighbouring subunit and makes both van der Waals interactions and hydrogen bonds with neighbouring E (Glu228) side chains (Fig. 4c). The W residue is particularly sensitive to mutation; Ca^{2+} uptake was not observed for any of the mutants tested (alanine, phenylalanine or tyrosine; Fig. 4d).

Ca^{2+} binding

Roughly spherical density is located at three sites in the upper portion of the pore (Fig. 4a). The sites are evenly spaced approximately 6 Å apart and located on the symmetry axis. Densities in site 1 and site 2 are within the selectivity filter and can be ascribed to Ca^{2+} ions on the basis of their chemical environments. Density in site 3, which is located in the upper portion of the vestibule, is harder to assign as it is approximately 6 Å away from the surrounding Thr231 and Tyr232 residues (measured to atom centres) and may represent Ca^{2+} or another small molecule. Site 1 is positioned in the centre of the ring of Asp225 residues. The distance between site 1 and each Asp225 carboxylate is approximately 4 Å (between atom centres), which suggests water-mediated interactions with these side chains. Site 2 is located within the ring of Glu228 residues and only about 2.3 Å from their carboxylates, indicating direct coordination.

The surface of the channel facing the intermembrane space is strongly electronegative (Extended Data Fig. 7e). In addition to the ring of Asp225 residues surrounding site 1, Asp221 (Glu257 in human MCU) is present on each turret (Fig. 3). The preponderance of electronegativity would tend to concentrate cations at the entrance to the pore. We envision the following scenario. A Ca^{2+} ion reaching the pore would remain hydrated in site 1. As it moves inward towards site 2 it would shed water to directly interact with the ring of Glu228 residues, perhaps retaining axial water molecules above and below it. On the basis of this direct coordination, and as site 2 has the strongest density in the entire cryo-EM map (Fig. 4a), we suspect that site 2 represents the high-affinity binding site for Ca^{2+} in MCU. After emerging from the selectivity filter, a Ca^{2+} ion could become more hydrated in the vestibule (for example, in site 3). We hypothesize that ion selectivity

© 2018 Springer Nature Limited. All rights reserved.

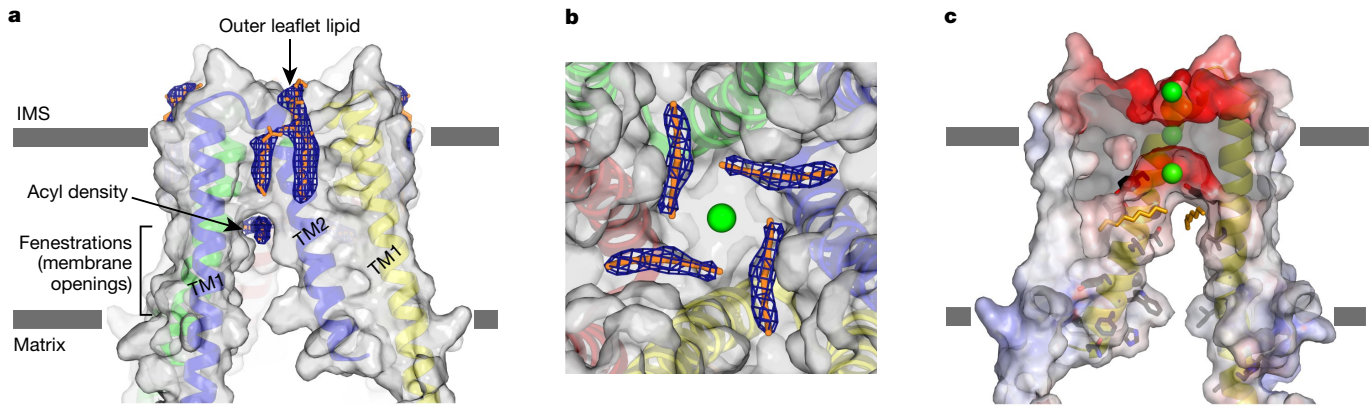


Fig. 5 | Membrane interactions. a, Ordered lipids within the transmembrane region. Densities adacent to each subunit (blue mesh, 8σ contour) are ascribed to a lipid of the outer leaflet and an acyl chain within a lateral membrane opening (fenestration). The partially transparent molecular surface (grey) reveals a ribbon representation of the channel, in which each subunit is coloured uniquely. **b**, Close-up view, orthogonal to **a**, from within the matrix, showing densities corresponding to the four acyl chains within the fenestrations (site 3 is a green sphere). **c**, The vestibule is hydrophobic. A slice of the molecular surface shown from **a** (the portion facing the viewer is removed) is coloured according to electrostatic potential: light grey regions are neutral; red, -8 kT e^{-1}; blue, $+8$ kT e^{-1}. Green spheres indicate sites 1–3. One subunit is shown as yellow ribbons with amino acids facing the vestibule drawn as sticks: Ile198, Ala199, Gly202, Leu206 and Tyr209 from TM1 and Tyr232, Gly235, Leu236, Thr238, Leu239, Tyr243 and Phe246 from TM2; the corresponding amino acids in human MCU are: Val235, Leu236, Gly239, Met243, Gln246, Tyr268, Thr271, Tyr272, Ser274, Ala275, Tyr279 and Phe282, respectively. Basic residues at the matrix boundary of the membrane (for example, Arg193, Arg197 and Arg251 of CeMCU) may interact with phospholipid headgroups.

for Ca^{2+} involves a knock-off mechanism between sites 1 and 2. In this mechanism, a Ca^{2+} ion moving inward from site 1 would displace the Ca^{2+} ion in site 2 into the vestibule, and site 1 would be refilled from the intermembrane space. Such a mechanism, involving multiple ions in single file, can allow rapid ion throughput in the context of high selectivity[28].

Site 2 may underlie another characteristic of MCU. Unlike many Ca^{2+} channels that efficiently conduct both Na^+ and K^+ when Ca^{2+} and other divalent cations are removed, MCU efficiently conducts Na^+ but not K^+ under these conditions[8]. The dimensions of the selectivity filter at site 2 may explain this electrophysiological hallmark. We hypothesize that Na^+ could pass through site 2 when it is not occupied by Ca^{2+} because Na^+ and Ca^{2+} have nearly the same ionic radius (1.02 Å and 1.00 Å, respectively). On the other hand, K^+ has an ionic radius of 1.38 Å and would be too large to fit in site 2. Correspondingly, the selectivity filter in K^+ channels is slightly wider than in MCU. In the K^+ channel KcsA[29], for example, the oxygen atoms of the selectivity filter that coordinate K^+ are separated by 4.7 Å across the pore in comparison to approximately 3.9 Å separation in MCU (measured between atom centres). This discrimination on the basis of size is an indication that the selectivity filter of MCU is relatively rigid at site 2. We postulate that site 2 preferentially binds Ca^{2+} over Na^+ because its $+2$ charge is better matched to direct coordination by the glutamate residues, and that site 2 would be constitutively occupied by Ca^{2+} under cellular conditions. When sites 1 and 2 are occupied, the local concentration of Ca^{2+} in the selectivity filter would be approximately 8 M (2 ions in a volume of approximately 4×10^{-25} l), which is more than 1 million times the typical concentration of Ca^{2+} in the cytosol. Selectivity for Ca^{2+} in MCU appears to be a combination of attractive, repulsive and excluded-volume forces.

The N-terminal domains

Each individual NTD of CeMCU adopts the same structure (Cα root mean square deviation (r.m.s.d.) = 0.5 Å), and superposes with X-ray structures of isolated NTDs of human MCU (Cα r.m.s.d. = 2 Å)[30,31], despite 16% sequence identity (Extended Data Fig. 9a). Inspection of the crystal lattice from the X-ray structures reveals an arrangement of NTDs that is similar to the crescent we observe in the structure of the full-length zebrafish channel (Extended Data Fig. 9b, c). This correspondence suggests that the NTDs of human MCU adopt a similar organization in the assembled channel. We find that, like human MCU[18,30], zebrafish MCU can catalyse Ca^{2+} uptake without an NTD (Extended Data Fig. 2a–c). Nevertheless, it is possible that the NTDs have regulatory properties, which may include modulation by matrix cations[31] and binding of MCUR1[30].

Membrane interactions

Densities ascribed (on the basis of shape and position) to four lipid molecules per channel are located on the membrane-exposed surfaces adjacent to the TM2 helices (Fig. 5a). Although the type of lipid cannot be discerned from the density, the lipids would belong to the outer leaflet and are oriented in a conventional manner with their head groups facing the intermembrane space.

Within the inner leaflet, TM1 and TM2 of the same subunit are separated by a few Ångstroms and this creates lateral membrane openings ('fenestrations') in the transmembrane molecular surface that would expose the vestibule to the membrane (Fig. 5a). Four horizontally oriented tubes of density consistent with acyl chains of lipid molecules are observed within the fenestrations just below site 3 (Fig. 5a, b). The densities are positioned similarly to acyl chains observed in fenestrations of certain other cation channels, including two-pore domain potassium channels and a voltage-dependent sodium channel[21,32–34]. A minor conformational change in the transmembrane helices, such as a hinge-like motion at or near Pro299 or Gly235 on TM2 (Fig. 3), could seal the fenestrations, raising hypotheses regarding possible ion permeation and gating mechanisms and the potential involvement of lipids in them.

Discussion

The presence of lipids in the vestibule might prevent ion permeation, and thus a wider lower portion of the channel, as observed in the structure of CeMCU, may represent a non-conductive ('closed') conformation of the channel. On the other hand, with its wide vestibule, the structure may represent a conductive conformation. If this is the case, perhaps lipids are involved in the inward rectification exhibited by MCU, wherein the channel preferentially catalyses the flow of Ca^{2+} towards the matrix[8]. In this mechanism, Ca^{2+} ions emerging from the selectivity filter into site 3 would be pulled inward by the voltage gradient across the membrane and could transiently displace lipids along their path through the vestibule. Ca^{2+} ions would not have sufficient

© 2018 Springer Nature Limited. All rights reserved.

potential energy to efficiently pass through the lipids without the negative voltage of the matrix. In the absence of inward ion flow, the vestibule could fill with lipids and MCU would present itself to the matrix as collection of transmembrane helices with intervening lipids rather than an aqueous pore. These lipids 'covering up' the selectivity filter would tend to prevent outward ion flow and this would be manifest to the experimentalist as inward rectification when applying a positive voltage inside the matrix. The legs of the channel appear suited to allow lipid movement. Entire lipid molecules of the inner leaflet, along with their head groups, could diffuse in and out of the vestibule between the legs. In line with this hypothesis, the surfaces of TM1 and TM2 lining the vestibule are hydrophobic and basic residues on the legs appear poised to interact with phospholipid head groups (Fig. 5c). Analogous residues are present in human MCU. The structure establishes a foundation to probe lipid involvement.

The structures of zebrafish MCU and CeMCU reveal a new architecture for a Ca^{2+} channel. The pore of MCU differs from the large superfamily of cation channels that share a pore architecture that was first observed for the K^+ channel KcsA[35]. Unlike those channels, which have loop-like secondary structures in their selectivity filters, the selectivity filter of MCU is formed by α-helices that coordinate Ca^{2+} with rings of acidic amino acids. The structures suggest that MCU combines a multi-ion selectivity filter, a discriminatory Ca^{2+} binding site, an electronegative surface exposed to the intermembrane space, and a vestibule to lower energy barriers for Ca^{2+} influx into mitochondria and achieve high selectivity for Ca^{2+}. They also indicate potential sites of channel regulation, not only by EMRE, MICU1 and MICU2 in the membrane and intermembrane space but also by other functionalities within the matrix.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0331-8.

Received: 4 March 2018; Accepted: 19 June 2018;
Published online 11 July 2018.

1. Deluca, H. F. & Engstrom, G. W. Calcium uptake by rat kidney mitochondria. *Proc. Natl Acad. Sci. USA* **47**, 1744–1750 (1961).
2. Montero, M. et al. Chromaffin-cell stimulation triggers fast millimolar mitochondrial Ca^{2+} transients that modulate secretion. *Nat. Cell Biol.* **2**, 57–61 (2000).
3. Baughman, J. M. et al. Integrative genomics identifies MCU as an essential component of the mitochondrial calcium uniporter. *Nature* **476**, 341–345 (2011).
4. De Stefani, D., Raffaello, A., Teardo, E., Szabò, I. & Rizzuto, R. A forty-kilodalton protein of the inner membrane is the mitochondrial calcium uniporter. *Nature* **476**, 336–340 (2011).
5. Chaudhuri, D., Sancak, Y., Mootha, V. K. & Clapham, D. E. MCU encodes the pore conducting mitochondrial calcium currents. *eLife* **2**, e00704 (2013).
6. Kamer, K. J. & Mootha, V. K. The molecular era of the mitochondrial calcium uniporter. *Nat. Rev. Mol. Cell Biol.* **16**, 545–553 (2015).
7. Gunter, K. K. & Gunter, T. E. Transport of calcium by mitochondria. *J. Bioenerg. Biomembr.* **26**, 471–485 (1994).
8. Kirichok, Y., Krapivinsky, G. & Clapham, D. E. The mitochondrial calcium uniporter is a highly selective ion channel. *Nature* **427**, 360–364 (2004).
9. Bick, A. G., Calvo, S. E. & Mootha, V. K. Evolutionary diversity of the mitochondrial calcium uniporter. *Science* **336**, 886 (2012).
10. Sancak, Y. et al. EMRE is an essential component of the mitochondrial calcium uniporter complex. *Science* **342**, 1379–1382 (2013).
11. Perocchi, F. et al. MICU1 encodes a mitochondrial EF hand protein required for Ca^{2+} uptake. *Nature* **467**, 291–296 (2010).
12. Plovanich, M. et al. MICU2, a paralog of MICU1, resides within the mitochondrial uniporter complex to regulate calcium handling. *PLoS One* **8**, e55785 (2013).
13. Mallilankaraman, K. et al. MCUR1 is an essential component of mitochondrial Ca^{2+} uptake that regulates cellular metabolism. *Nat. Cell Biol.* **14**, 1336–1343 (2012).
14. Vais, H. et al. MCUR1, CCDC90A, is a regulator of the mitochondrial calcium uniporter. *Cell Metab.* **22**, 533–535 (2015).
15. Chaudhuri, D., Artiga, D. J., Abiria, S. A. & Clapham, D. E. Mitochondrial calcium uniporter regulator 1 (MCUR1) regulates the calcium threshold for the mitochondrial permeability transition. *Proc. Natl Acad. Sci. USA* **113**, E1872–E1880 (2016).
16. Kovács-Bogdán, E. et al. Reconstitution of the mitochondrial calcium uniporter in yeast. *Proc. Natl Acad. Sci. USA* **111**, 8985–8990 (2014).
17. Tsai, M.-F. et al. Dual functions of a small regulatory subunit in the mitochondrial calcium uniporter complex. *eLife* **5**, e15545 (2016).
18. Oxenoid, K. et al. Architecture of the mitochondrial calcium uniporter. *Nature* **533**, 269–273 (2016).
19. Moraes, I., Evans, G., Sanchez-Weatherby, J., Newstead, S. & Stewart, P. D. Membrane protein structure determination — the next generation. *Biochim. Biophys. Acta* **1838** (1 Pt A), 78–87 (2014).
20. Long, S. B., Tao, X., Campbell, E. B. & MacKinnon, R. Atomic structure of a voltage-dependent K^+ channel in a lipid membrane-like environment. *Nature* **450**, 376–382 (2007).
21. Payandeh, J., Scheuer, T., Zheng, N. & Catterall, W. A. The crystal structure of a voltage-gated sodium channel. *Nature* **475**, 353–358 (2011).
22. Wu, J. et al. Structure of the voltage-gated calcium channel Cav1.1 complex. *Science* **350**, aad2395 (2015).
23. Shen, H. et al. Structure of a eukaryotic voltage-gated sodium channel at near-atomic resolution. *Science* **355**, eaal4326 (2017).
24. Yan, Z. et al. Structure of the rabbit ryanodine receptor RyR1 at near-atomic resolution. *Nature* **517**, 50–55 (2015).
25. Zalk, R. et al. Structure of a mammalian ryanodine receptor. *Nature* **517**, 44–49 (2015).
26. Fan, G. et al. Gating machinery of InsP3R channels revealed by electron cryomicroscopy. *Nature* **527**, 336–341 (2015).
27. Liao, M., Cao, E., Julius, D. & Cheng, Y. Structure of the TRPV1 ion channel determined by electron cryo-microscopy. *Nature* **504**, 107–112 (2013).
28. Hille, B. *Ionic Channels of Excitable Membranes* 3rd edn (Sinauer Associates, Sunderland, 2001).
29. Zhou, Y., Morais-Cabral, J. H., Kaufman, A. & MacKinnon, R. Chemistry of ion coordination and hydration revealed by a K^+ channel–Fab complex at 2.0 A resolution. *Nature* **414**, 43–48 (2001).
30. Lee, Y. et al. Structure and function of the N-terminal domain of the human mitochondrial calcium uniporter. *EMBO Rep.* **16**, 1318–1333 (2015).
31. Lee, S. K. et al. Structural insights into mitochondrial calcium uniporter regulation by divalent cations. *Cell Chem. Biol.* **23**, 1157–1169 (2016).
32. Miller, A. N. & Long, S. B. Crystal structure of the human two-pore domain potassium channel K2P1. *Science* **335**, 432–436 (2012).
33. Brohawn, S. G., Campbell, E. B. & MacKinnon, R. Physical mechanism for gating and mechanosensitivity of the human TRAAK K^+ channel. *Nature* **516**, 126–130 (2014).
34. Dong, Y. Y. et al. K2P channel gating mechanisms revealed by structures of TREK-2 and a complex with Prozac. *Science* **347**, 1256–1259 (2015).
35. Doyle, D. A. et al. The structure of the potassium channel: molecular basis of K^+ conduction and selectivity. *Science* **280**, 69–77 (1998).

Acknowledgements We thank N. Grigorieff, members of his laboratory, and the staff at the Howard Hughes Medical Institute Cryo-EM facility for training in cryo-EM; R. K. Hite and members of the Long laboratory for discussions; the staff of the New York Structural Biology Center Simons Electron Microscopy Center, M. Ebrahim, and M. J. de la Cruz of the Memorial Sloan Kettering Cancer Center Cryo-EM facility for help with data collection; and J. Goldberg for spectrophotometer use. This work was supported, in part, by an NIH core facilities grant to Memorial Sloan Kettering Cancer Center (P30 CA008748), by an NIH Medical Scientist Training Program grant (T32GM007739 for A.F.S.), and by an NIH grant (R01GM094273 to S.B.L.).

Reviewer information *Nature* thanks P. Yuan and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions R.B. and C.W. performed cryo-EM studies of CeMCU and zebrafish MCU, respectively, and other experiments. A.F.S. developed MCU and EMRE double-knockout cells. S.B.L. directed the research and assisted with cryo-EM. All authors contributed to data analysis and the preparation of the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0331-8.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0331-8.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to S.B.L.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Cloning, expression and purification of CeMCU. *Cyphellophora europaea* MCU (CeMCU; UniProt accession W2SDE2) was selected as a candidate for structural studies from among numerous MCU orthologues that were evaluated using fluorescence-detection size-exclusion chromatography (FSEC) with HEK-293 cells[36]. The cDNA of CeMCU (synthesized by BioBasic) was ligated into the EcoRI–SalI restriction sites of the *Pichia pastoris* expression vector pPICZ-C (Invitrogen Life Technologies), encoding the full-length protein followed by a C-terminal antibody-affinity tag (GEGEEF) that is recognized by the anti-tubulin antibody YL1/2[37]. Transformation into *P. pastoris* (SMD1168 strain, Invitrogen), expression and cryo-lysis were performed as previously described[38].

Mild detergents (n-dodecyl-β-D-maltoside and digitonin) that are commonly used for structural studies of membrane proteins[19] were used for preparation of CeMCU and zebrafish MCU. All purification steps were carried out at 4 °C. For CeMCU, lysed *P. pastoris* cells (~150 g) were re-suspended in 1.5 L of RB buffer containing 20 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES), pH 7.5, 150 mM NaCl, 1 mM $CaCl_2$, 0.15 mg/mL DNase I (Sigma-Aldrich), 1.5 μg/mL Leupeptin (Sigma-Aldrich), 1.5 μg/mL Pepstatin A (Sigma-Aldrich), 1 mM Benzamidine (Sigma-Aldrich), 1 mM 4-(2-aminoethyl) benzenesulfonyl fluoride hydrochloride (AEBSF, Gold Biotechnology), 1 mM phenylmethanesulfonyl fluoride (PMSF, Acros Organics) and 1:500 dilution of Aprotonin (Sigma-Aldrich). Cell debris was removed by centrifugation (3 × 7000*g* for 20 min), and cellular membranes were collected by centrifugation at 43,000*g* for 20 min. The pelleted membranes were resuspended in 100 mL of RB buffer supplemented with 20% (v/v) glycerol, flash-frozen in liquid nitrogen, stored at −80 °C. Membranes (~33 mL, corresponding to ~40 g of lysed cells) were thawed, diluted with equal volume of RB buffer, and solubilized with 2 g of n-dodecyl-β-maltoside (DDM, Anatrace) by rotating for 1 h at 4 °C. The sample was then centrifuged at 43,000*g* for 45 min and the supernatant was filtered through a 0.22-μm polystyrene membrane (Millipore). YL1/2 antibody (IgG, expressed from hybridoma cells and purified by ion-exchange chromatography using standard methods) was coupled to CNBr-activated Sepharose beads (GE Healthcare) according to manufacturer's protocol. Approximately 0.4 mL of YL1/2 antibody beads were added to the sample for each 1 g of *P. pastoris* cells and the mixture was rotated at 4 °C for 1 h. Antibody beads were collected on a column support, washed with ~5 column volumes of a buffer containing 20 mM HEPES, pH 7.5, 150 mM NaCl, 1 mM $CaCl_2$, 2 mM DDM and 0.1 mg/mL 1',3'-bis[1,2-dioleoyl-*sn*-glycero-3-phospho]-*sn*-glycerol (18:1 cardiolipin, Avanti Polar Lipids), and the protein was eluted with a buffer containing 100 mM HEPES, pH 7.5, 150 mM NaCl, 1 mM $CaCl_2$, 2 mM DDM, 0.1 mg/mL cardiolipin and 1 mM Glu-Glu-Phe peptide (Peptide 2.0). The elution fraction was concentrated to ~5 mg/mL using a 100 kDa concentrator (Amicon Ultra; EMD Millipore) and run on a Superose-6 10/300 Increase size-exclusion chromatography (SEC) GL column, equilibrated with a buffer containing 20 mM HEPES pH 7.5, 150 mM NaCl, 1 mM $CaCl_2$, 0.5 mM digitonin (Cayman Chemical Company) and 0.05 mg/mL cardiolipin. (Digitonin was used at the end of the purification procedure because it produced more uniform ice on the cryo-EM grids than DDM.) The elution fractions containing the CeMCU protein were pooled and concentrated to ~5 mg/mL using a 100 kDa concentrator (Vivaspin-2). Purified protein was used immediately for cryo-Electron Microscopy (cryo-EM) grid preparation.

Cloning, expression and purification of zebrafish MCU. *Danio rerio* (zebrafish) MCU was selected as a candidate for structural studies on the basis of good protein expression in HEK293S GnTI $^-$ cells (ATCC, not tested for mycoplasma contamination), the ability to catalyse mitochondrial calcium uptake when co-expressed with EMRE, and good biochemical stability in detergent-containing solutions. The cDNA of zebrafish MCU (UniProt accession zgc:153607, synthesized by Synbio Tech Inc.) was ligated into the XhoI–EcoRI restriction sites of a BacMam expression vector[39], encoding the full-length protein followed by a C-terminal 1D4 tag (TETSQVAPA) that is recognized by the Rho-1D4 monoclonal antibody[40]. The resulting plasmid was transformed to DH10Bac *E. coli* competent cells to generate bacmid. Recombinant P1 baculoviruses were generated in Sf9 cells (Invitrogen, not tested for mycoplasma contamination) cultured in sf-900 III SFM medium (GIBCO) using cellfectin II reagents (Invitrogen). For protein expression, 100 mL P3 virus was used to infect 1 L culture of HEK293S GnTI $^-$ cells (3 × 10^6 cells/mL) in Freestyle 293 media (GIBCO) supplemented with 2% FBS (GIBCO, catalogue A31606-02). After infection, the cells were incubated at 37 °C for 16 h, and then 5 mM sodium butyrate (Sigma-Aldrich) was added to the cell culture and cells were cultured at 30 °C for another 72 h before harvest.

2 L of cell pellet was resuspended in 100 mL lysis buffer (40 mM HEPES pH 7.4 150 mM NaCl, 5 μM $CaCl_2$, 0.15 mg/ml DNase I, 1.5 μg/ml Leupeptin, 1.5 μg/ml Pepstatin A, 1 mM AEBSF, 1 mM Benzamidine, 1 mM PMSF and 1:500 dilution of Aprotinin), and then solubilized by adding lauryl maltose neopentyl glycol (LMNG, Anatrace, from a 10% stock in water) to a final concentration of 1% and stirred at 4 °C for 1.5 h. Detergent-solubilized material was separated from the insoluble fraction by centrifugation at 43,000*g* for 1 h and the supernatant was filtered through a 0.22-μm polystyrene membrane (Millipore). Rho-1D4 monoclonal antibody was coupled to CNBR-activated Sepharose beads (GE healthcare) according to the manufacturer's protocol. 5 mL Rho-1D4 monoclonal antibody beads were added to the sample and rotated at 4 °C for 2 h. The beads were washed with 100 mL of buffer A (20 mM HEPES pH 7.4 150 mM NaCl, 5 μM $CaCl_2$, 0.5 mM digitonin and 0.01 mg/ml 18:1 cardiolipin). Zebrafish MCU protein was eluted by incubating with buffer A containing 2 mg/ml 1D4 peptide (TETSQVAPA, synthesized by Peptide 2.0) overnight at 4 °C, and further purified by size-exclusion chromatography on a Superose 6 increase, 10/300 GL column (GE Healthcare) equilibrated with buffer B (20 mM Tris(hydroxymethyl)-aminomethane (Tris) pH 8.5, 150 mM NaCl, 5 μM $CaCl_2$, 0.5 mM digitonin, 0.01 mg/ml cardiolipin and 0.4 mM Tris(2-carboxyethyl)phosphine). The peak fractions were pooled, concentrated to 5 mg/ml using a 100 kDa concentrator (Vivaspin-2), and immediately used for cryo-EM grid preparation.

EM sample preparation and data acquisition. 5 μL of 5 mg/mL freshly purified CeMCU and zebrafish MCU protein samples were loaded into glow-discharged (10 s) Quantifoil R1.2/1.3 holey carbon grids (Au 400, Electron Microscopy Sciences) and plunge-frozen in liquid nitrogen-cooled liquid ethane, using a Vitrobot Mark IV (FEI) operated at room temperature with a blotting time of 2 s under blot force of 0 at 100% humidity. Grids were clipped and loaded into a 300 keV Titan Krios microscope (FEI) equipped with a Gatan K2 Summit direct electron detector (Gatan). Grids were screened on the bases of particle distribution and density, and images from the best ones were recorded. The CeMCU micrographs were collected at a magnification of 29,000 × with a super-resolution pixel size of 0.4307 Å and a defocus range of −0.9 to −2.9 μm. The dose rate was 8 electrons per physical pixel per second, and images were recorded for 8 s with 0.2 s subframes (40 total frames) corresponding to a total dose of approximately 80 electrons per $Å^2$. The micrographs for zebrafish MCU were collected at a magnification of 37,000 × with a super-resolution pixel size of 0.346 Å and a defocus range of −1.2 to −3.2 μm. The dose rate was 8 electrons per physical pixel per second and images were recorded for 6 s with 0.13 s subframes (46 total frames) corresponding to a total dose of approximately 100 electrons per $Å^2$.

Image processing of CeMCU. Extended Data Figure 3 shows the cryo-EM workflow of CeMCU. Movie stacks were gain-corrected and twofold Fourier cropped to a calibrated pixel size of 0.8613 Å, and motion corrected and dose weighted using MotionCor2[41]. Contrast Transfer Function (CTF) estimates for motion-corrected micrographs were performed in CTFFIND4 using frames 1-40[42]. Subsequent image processing was carried out with RELION 2.1[43], using a particle box size of 384 pixels and a mask diameter of 160 Å, unless stated otherwise. The initial data set (~3700 movies) was obtained from two equivalent data collection sessions. Each micrograph was inspected manually; poor-quality ones and those having CTF estimation fits to lower than 5 Å were discarded. Approximately 2000 particles were picked manually and these were used for reference-free 2D classification to generate templates for automatic particle picking. Auto-picking (using 8 of the representative 2D classes containing different orientations as templates) from ~3,700 micrographs yielded 240,516 particles, which were further cleaned-up by removing outlier particles using two rounds of reference-free 2D classification with 100 classes (for example, to remove contaminants or empty micelles), and resulted in 201,281 particles. An initial 3D model from these particles was generated without imposed symmetry (C1 symmetry) with RELION 2.1. Using this initial model, 3D classification with C1 symmetry was performed, sorting the particles into 10 classes. The 10 classes had similar shapes and all indicated a tetrameric architecture. By visual inspection in UCSF Chimera[44], 8 "good" classes (representing 171,407 particles) were combined and subjected to 3D refinement. This yielded a reconstruction at 4.8 Å overall resolution in C1 symmetry. "Polished" particle images, which are corrected for individual particle movements, were generated in Relion using aligned movie frames that were output from MotionCor2. 3D refinement using the 171,407 polished particles and a mask that removes the micelle region yielded a reconstruction to 4.5 Å overall resolution with C1 symmetry. Well-defined density for α-helices enabled determination of the appropriate hand for the reconstruction at this point (Extended Data Fig. 3). The C2 overall symmetry of the molecule was evident from this map: it has fourfold symmetry within the TMD and twofold symmetry within the matrix region (Fig. 1a). Refinement using C2 symmetry in RELION 2.1 improved the overall resolution to 4.1 Å. All resolution estimates are based on gold-standard Fourier shell correlation (FSC) calculations.

To improve the reconstruction further, an additional ~2340 micrographs were collected using the same microscope and the same data collection strategy. These were motion-corrected using MotionCor2 as indicated above. Auto-picked particles were subjected to one round of reference-free 2D classification and merged with the previous data set at the "un-polished" stage to yield a total of 547,637 particles. 3D refinement in RELION 2.1 was performed using all of these particles (with C2 symmetry) and the particles were "polished" together as a single data set using methods described above. The polished particles were then subjected to 3D classification, with C2 symmetry, sorting them into 10 classes. 8 good classes

© 2018 Springer Nature Limited. All rights reserved.

containing 376,541 polished particles were combined and subjected to 3D refinement in RELION 2.1 but this did not improve the overall resolution beyond 4.1 Å. All of these particles were imported into the cisTEM cryo-EM software package for further refinement[45]. 3D refinement in cisTEM, using the previously used mask, but where the micelle region outside the mask was low-pass filtered at 15 Å rather than removed, yielded the final reconstruction at 3.2 Å overall resolution. We suspect that this low-pass filter and the weighting scheme used for reconstruction in cisTEM were the primary reasons for the improved resolution[45]. The map was B-factor sharpened using a B-factor of -90 Å^2 and low-pass filtered at 3.2 Å resolution (using cisTEM). Independent half-maps were generated and only one of them was used to build and refine the atomic model; the other was used only for validation. We note that the additional data set was important to extend the resolution of the reconstruction; refinement in cisTEM using only the initial 171,407 polished particles yielded a reconstruction to 3.8 Å overall resolution. Estimation of the local resolution of the map was done using blocres[46]. The maps of CeMCU shown in Figures are combined maps, which were sharpened (using a *B*-factor of -90 Å^2) and low-pass filtered at 3.2 Å using cisTEM.

Model building and refinement. The atomic model of CeMCU was manually built into one of the half-maps (sharpened using a B-factor of -90 Å^2 and low-pass filtered at 3.2 Å resolution) and refined in real space using the COOT software[47]. The crystal structure of the NTD from human MCU (PDB: 5KUI) was used as a reference for the corresponding regions. Further refinement of the atomic model was carried out in PHENIX[48] using real-space refinement against the same half-map. The final model has good stereochemistry and good Fourier shell correlation with the other half-map as well as the combined map (Extended Data Fig. 6e and Extended Data Table 1). Structural Figures were prepared with Pymol (pymol.org), Chimera[44], and HOLE[49]. Electrostatic calculations used the APBS[50] plugin in Pymol.

Image processing of zebrafish MCU. Extended Data Figure 4 shows the cryo-EM workflow for zebrafish MCU. Movie stacks were gain-corrected, twofold Fourier cropped to a pixel size of 0.692 Å, motion corrected, and dose weighted using MotionCor2[3]. Contrast transfer function (CTF) estimates were performed in CTFFIND4 using frames 1-46[4]. Subsequent image processing was carried out in RELION 2.1[5]. Generating an improved initial model was crucial for 3D refinement of zebrafish MCU (Extended Data Fig. 4a). The improved initial model was generated using a subset of particles but refinement (Extended Data Fig. 4b) was done using the majority of particles as described below. Approximately 3000 particles were manually picked and subjected to 2D classification. Nine 2D-class averages were selected as templates for autopicking of 269,574 particles from ~3500 micrographs in RELION 2.1. After 2D classification to remove a few bad classes, 203,213 particles were used for ab-initio 3D model generation (with 4 classes) in cryoSPARC[6]. The best 3D reconstruction and the associated particle images were used for further 3D classification in RELION 2.1, sorting the particles into 6 classes. Three "good" classes (91,727 particles) were combined and subjected to further 3D classification (using 6 classes). The best class (representing 15,775 particles) was improved by 3D refinement in RELION 2.1 and yielded the improved initial model (Extended Data Fig. 4a). At this point, particles from micrographs with CTF fits better than 5 Å were selected (196,272 particles) and used for homogeneous refinement in cryoSPARC using the improved initial model (Extended Data Fig. 4b). This yielded a reconstruction at 9.1 Å overall resolution. All 196,272 particles were imported into cisTEM for further refinement[8]. 3D refinement in cisTEM, using a mask to apply a 30 Å low-pass filter to the detergent micelle, improved the overall resolution to 8.5 Å. The map shown in Fig. 2 was B-factor sharpened using a B-factor of -90 Å^2 and low-pass filtered at 8.5 Å resolution (using cisTEM).

Crosslinking. For crosslinking of purified CeMCU in detergent, the protein was diluted to 0.2 mg/mL in buffer consisting of 20 mM HEPES, pH7.5, 150 mM NaCl, 1 mM $CaCl_2$, 1 mM DDM and 0.05 mg/mL cardiolipin. 0 to 1 mM concentrations of bis(sulfosuccinimidyl) suberate crosslinker (BS-3, Sigma-Aldrich, from a 17.5 mM stock in water) were added and the samples were incubated for 30 min at room temperature. The reactions were stopped by the addition of 140 mM Tris, pH 8.5, and analysed by 4–12% SDS-polyacrylamide gel electrophoresis under reducing conditions (100 mM dithiothreitol) using coomassie staining. For crosslinking of purified zebrafish MCU in detergent, the protein was diluted to 0.5 mg/mL in buffer consisting of 20 mM HEPES, pH7.5, 150 mM NaCl, 1 mM $CaCl_2$, 0.5 mM digitonin and 0.01 mg/mL cardiolipin. 0 to 1 mM concentrations of BS-3 crosslinker were added and the samples were incubated for 60 min on ice. The reactions were stopped by the addition of 100 mM Tris, pH 8.5, and analysed by 12% SDS-polyacrylamide gel electrophoresis under reducing conditions (100 mM dithiothreitol) using coomassie staining.

For crosslinking of human (Hs) MCU in MCU/EMRE knockout cell membranes, a chimaeric protein, consisting of the mitochondrial targeting sequence from *D. discoideum* MCU, fused to the mature form of human MCU (DdMTS–HsMCU) as previously used[16], followed by cDNA encoding a C-terminal Rho-1D4 affinity tag (amino acids: TETSQVAPA) and a stop codon, was cloned into the XhoI–EcoRI restriction sites of expression vector pCGFP-EU[36], transfected into MCU/EMRE knockout cells using Lipofectamine 3000 (Invitrogen), and the cells were cultured for 48 h at 37 °C with 5% CO_2. Cells were harvested and re-suspended in a buffer containing 50 mM HEPES, pH 7.5, 200 mM NaCl, 1 mM $CaCl_2$ and 1:500 dilution of Protease Inhibitor Cocktail III, EDTA-free (CalBiochem), and lysed by dounce homogenization and brief sonication (30 s) using a bath sonicator. Cell debris was removed by centrifugation (2500*g*, 4 °C, 15 min) and the membrane fractions were collected by centrifugation (100,000*g*, 4 °C, 1 h). The membrane fraction pellets were re-suspended in the same buffer using a dounce homogenizer followed by brief sonication (30 s), and treated with 0–575 μM of the membrane permeable crosslinker disuccinimidyl glutarate (DSG, Pierce, from a 30 mM stock in DMSO) for 30 min at room temperature. The reactions were stopped by the addition of 140 mM Tris, pH 8.5, run on a 4–12% SDS-polyacrylamide gel under the reducing conditions (100 mM dithiothreitol), and detected by western blot using the Rho-1D4 antibody[40].

Generation of MCU/EMRE knockout HEK293T cells. *SMDT1* and *MCU*, which encode human EMRE and MCU, respectively, were simultaneously targeted in HEK293T cells (Invitrogen, not tested for mycoplasma contamination) following a well-established CRISPR–Cas9 protocol[51]. In brief, sgRNA sequences targeting coding regions of *SMDT1* and *MCU* were selected using an online design tool (http://crispr.mit.edu/). Oligonucleotides incorporating target sequences for *SMDT1* (5′-caccgATGGCGATGTCTCCGCCGCA-3′, 5′-aaacTGCGGCGGAGACATCGCCATc-3′) and *MCU* (5′-caccgTGACAGCGTTCACGCCGGGA-3′, 5′-aaacTCCCGGCGTGAACGCTGTCAc-3′) were synthesized by IDT, annealed and cloned into the pSpCas9(BB)-2A-GFP vector (Addgene plasmid # 48138). Correctly inserted sgRNA sequences were verified for the *SMDT1*- and *MCU*-targeting vectors by Sanger sequencing (Genewiz). HEK293T cells were detached with trypsin and seeded into a 24-well plate with approximately 1.3×10^5 cells per well. Twenty-four hours after seeding, cells were transfected with 0.25 μg of each vector using Lipofectamine2000 (Invitrogen). Twenty-four hours post-transfection, cells were detached and single GFP-positive cells were sorted into individual wells of a 96-well plate containing DMEM (supplemented with 20% FBS, 100 U/ml penicillin, and 100 μg/ml streptomycin) using a BD FACSAria II cell sorter. After sorting, cells were incubated for 2 weeks, during which time media was changed every 3-5 days, and wells were monitored for the appearance of well-defined single colonies. Colonies were then transferred to a 24-well plate for subsequent outgrowth, DNA analysis, and functional characterization. Genomic DNA from individual colonies was prepared using QuickExtract solution according to the manufacturer's instructions. Loci containing the targeted sequences were PCR amplified and Sanger sequenced using the following primers: 5′-TAACGAATTCCACAGATAGCACGGGCCTTT-3′, 5′-TATGGTCGACTAGCGTATTCTTGGTCCGCC-3′ for *SMDT1* and 5′-TAACGAATTCACCCCCATGCAAGAATGGTA-3′, 5′-TATGGTCGACGTTCATCCTTGCTCATGGCATT-3′ for *MCU*. Putative double knockouts that were disrupted at both *SMDT1* and *MCU* loci (as judged by Sanger sequencing) were verified functionally using the mitochondrial Ca^{2+} uptake assay.

Mitochondrial Ca^{2+} uptake measurements. The cDNA for zebrafish MCU, zebrafish $MCU_{\Delta NTD}$ (deletion from residue 85 to residue 190), zebrafish EMRE (UniProt accession: Q08BI9), DdMTS–HsMCU chimaera, and human EMRE (UniProt accession: Q9H4I9) were sub-cloned into the pCGFP-EU mammalian expression vector with a C-terminal 1D4 tag (and without GFP)[36]. Mutations were generated using Q5 Site-Directed Mutagenesis Kit (New England BioLabs Inc.) and verified by DNA sequencing.

1.5 μg MCU plasmid and 1.5 μg EMRE plasmid were co-transfected into ~2×10^6 MCU/EMRE knockout cells using Lipofectamine2000 (Invitrogen). After transfection, the cells were cultured at 37 °C for ~48 h and then collected for calcium uptake assays and western blotting.

Calcium uptake assays of permeabilized MCU/EMRE knockout cells were performed according to a published protocol[16]. In brief, ~2×10^6 cells were resuspended in 1 mL of buffer containing 125 mM KCl, 2 mM K_2HPO_4, 1 mM $MgCl_2$, 10 μM ethylene glycol-bis(2-aminoethylether)-*N*,*N*,*N*′,*N*′-tetraacetic acid (EGTA, Gold Biotechnology), 5 mM malate, 5 mM glutamate, 5 mM succinate (Sigma-Aldrich) and 20 mM HEPES, pH 7.4, and loaded into a stirred quartz cuvette in a Hitachi F-2500 spectrophotometer (excitation: 506 nm, slit length: 3 nm, emission: 531 nm, slit length: 3 nm, sampling frequency: 2 Hz), with the temperature maintained at 22 °C by a circulating water bath. In a typical experiment, reagents were added into the cuvette in the following order: 1 μM cell-impermeable Calcium Green-5N (Life Technologies, from a 1 mM stock solution in water), 0.005% (w/v) digitonin (Sigma-Aldrich, from a 0.5% stock solution in water), 5 μM $CaCl_2$ (from a 1 mM stock solution in water), and either 1 μM ruthenium red (Sigma-Aldrich, from a 0.2 mM stock solution in water), 5 μM carbonyl cyanide *m*-chlorophenyl hydrazone (CCCP, Sigma-Aldrich, from a 1 mM stock solution in DMSO) or 10 μM ETH129 (Calcium ionophore II, Sigma-Aldrich, from a 2 mM stock solution in DMSO). For the data plotted in Fig. 4d, a rate of Ca^{2+} uptake was defined as the

© 2018 Springer Nature Limited. All rights reserved.

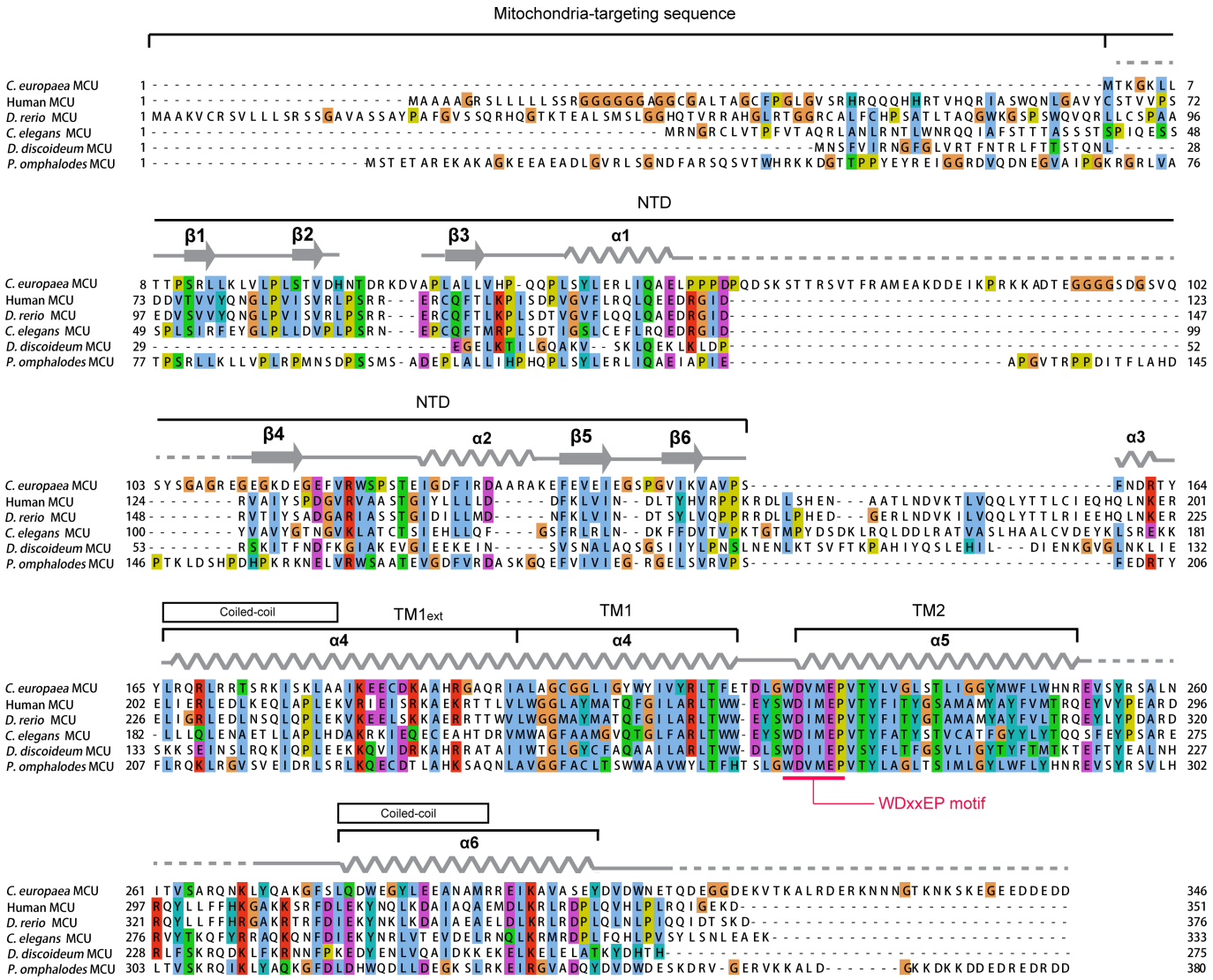
slope of a linear fit of the Ca^{2+} uptake curve between 240 and 260 s (for example, Extended Data Fig. 2d, f). Relative uptake (Fig. 4d) is the ratio of the slopes observed for mutant and wild type MCU channels.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. Atomic coordinates of CeMCU have been deposited in the PDB under accession 6DNF. The three-dimensional cryo-EM density maps have been deposited in Electron Microscopy Data Bank under accessions EMD-7971 (CeMCU) and EMD-7972 (zebrafish MCU).

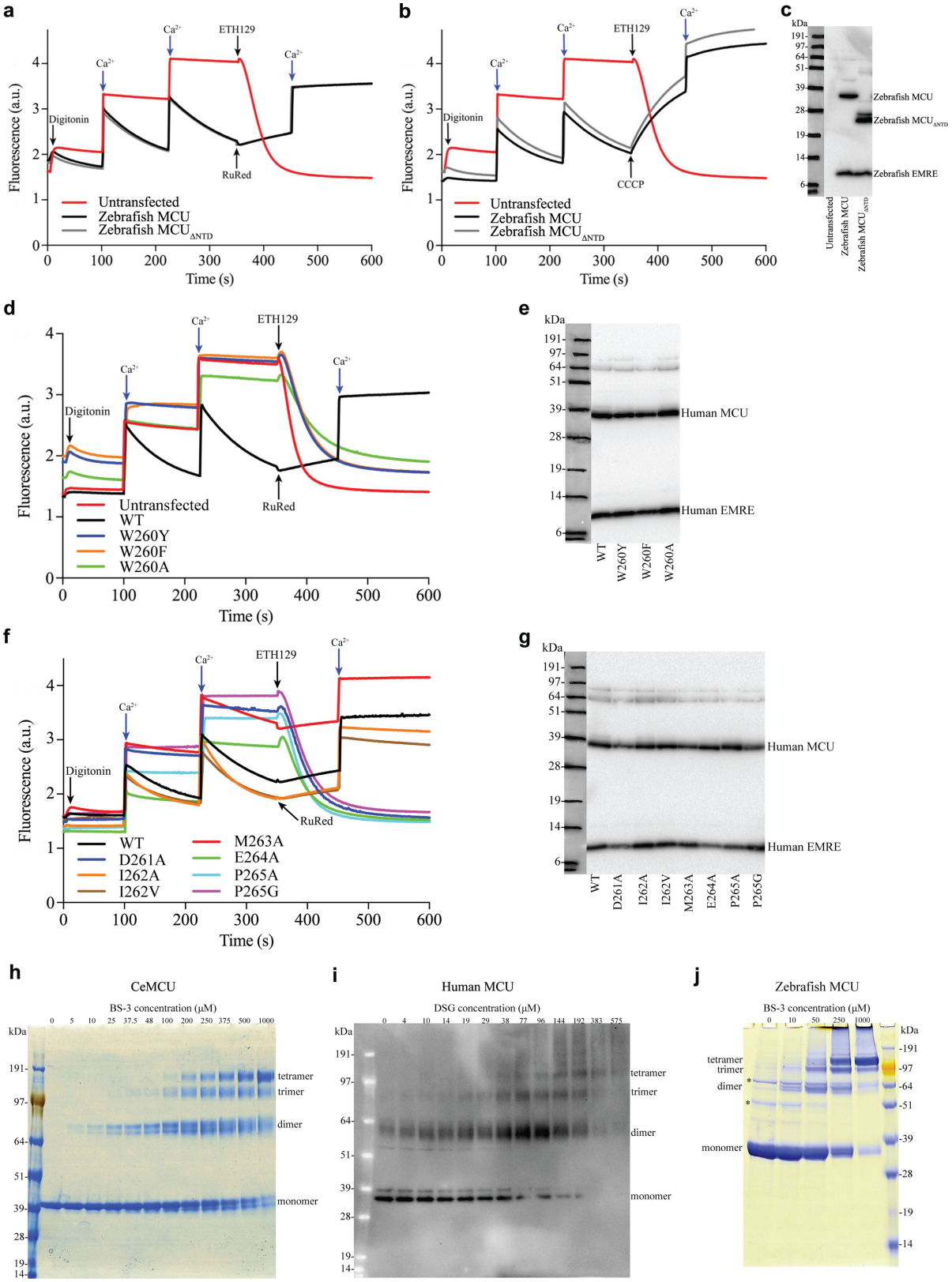
36. Kawate, T. & Gouaux, E. Fluorescence-detection size-exclusion chromatography for precrystallization screening of integral membrane proteins. *Structure* **14**, 673–681 (2006).
37. Kilmartin, J. V., Wright, B. & Milstein, C. Rat monoclonal antitubulin antibodies derived by using a new nonsecreting rat cell line. *J. Cell Biol.* **93**, 576–582 (1982).
38. Long, S. B., Campbell, E. B. & Mackinnon, R. Crystal structure of a mammalian voltage-dependent Shaker family K^+ channel. *Science* **309**, 897–903 (2005).
39. Goehring, A. et al. Screening and large-scale expression of membrane proteins in mammalian cells for structural studies. *Nat. Protocols* **9**, 2574–2585 (2014).
40. Molday, R. S. & MacKenzie, D. Monoclonal antibodies to rhodopsin: characterization, cross-reactivity, and application as structural probes. *Biochemistry* **22**, 653–660 (1983).
41. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
42. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).
43. Fernandez-Leiro, R. & Scheres, S. H. W. A pipeline approach to single-particle processing in RELION. *Acta Crystallogr. D Struct. Biol.* **73**, 496–502 (2017).
44. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
45. Grant, T., Rohou, A. & Grigorieff, N. cisTEM, user-friendly software for single-particle image processing. *eLife* **7**, e35383 (2018).
46. Cardone, G., Heymann, J. B. & Steven, A. C. One number does not fit all: mapping local variations in resolution in cryo-EM reconstructions. *J. Struct. Biol.* **184**, 226–236 (2013).
47. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 486–501 (2010).
48. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 213–221 (2010).
49. Smart, O. S., Neduvelil, J. G., Wang, X., Wallace, B. A. & Sansom, M. S. HOLE: a program for the analysis of the pore dimensions of ion channel structural models. *J. Mol. Graph.* **14**, 354–360 (1996).
50. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl Acad. Sci. USA* **98**, 10037–10041 (2001).
51. Ran, F. A. et al. Genome engineering using the CRISPR–Cas9 system. *Nat. Protocols* **8**, 2281–2308 (2013).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 1 | Structure-based sequence alignment. The amino acid sequences of *C. europaea*, *Pyronema omphalodes*, *D. discoideum*, *C. elegans*, zebrafish (*D. rerio*) and human MCUs are aligned and coloured according to the ClustalW convention (UniProt accession numbers: W2SDE2, U4LFM6, Q54LT0, Q21121, Q08BI9 and Q8NE86, respectively). The secondary structure is indicated with ribbons representing α-helices, solid lines representing structured loop regions, and dashed lines representing disordered regions. The WDXXEP signature sequence is highlighted with a red line.

© 2018 Springer Nature Limited. All rights reserved.

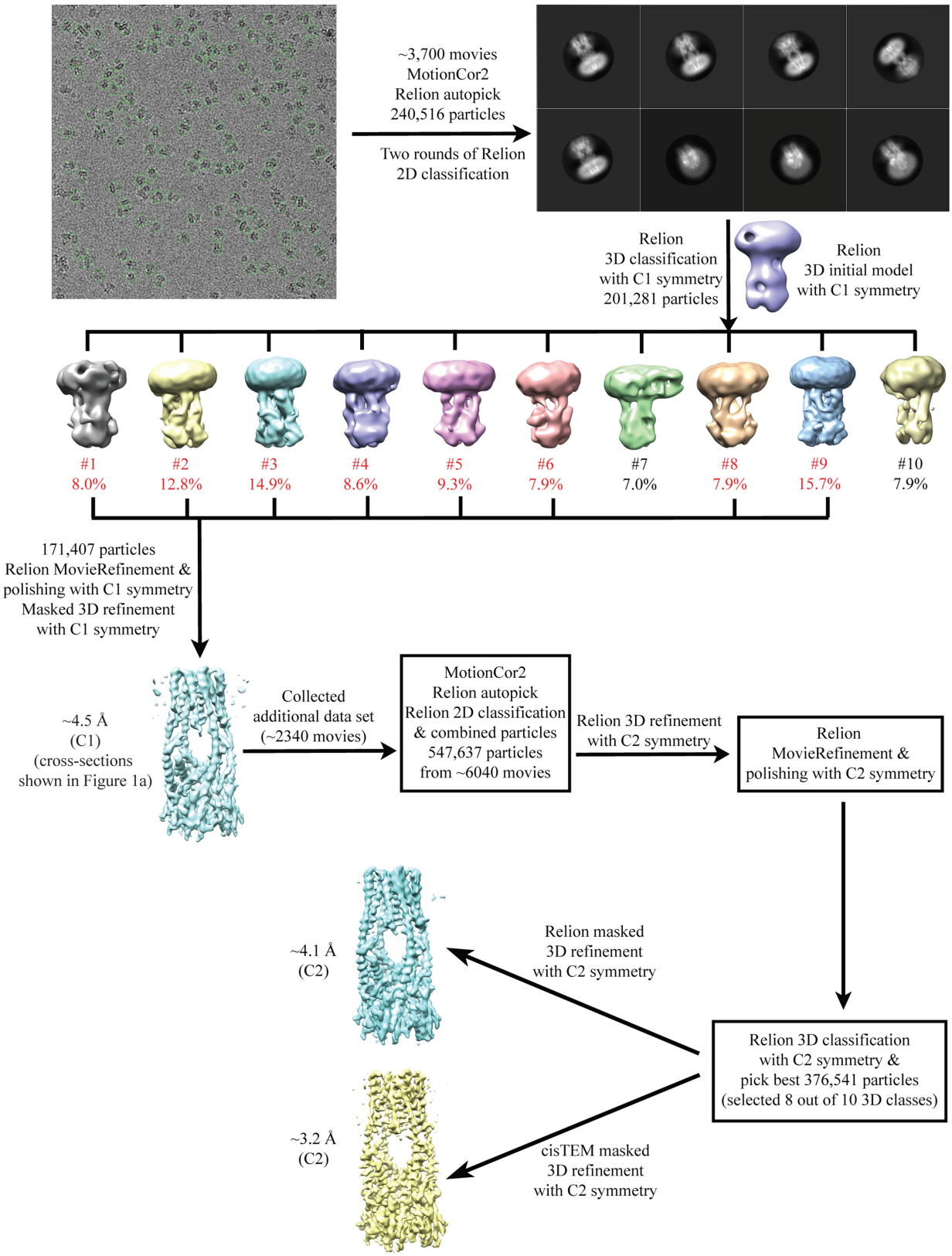


Extended Data Fig. 2 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

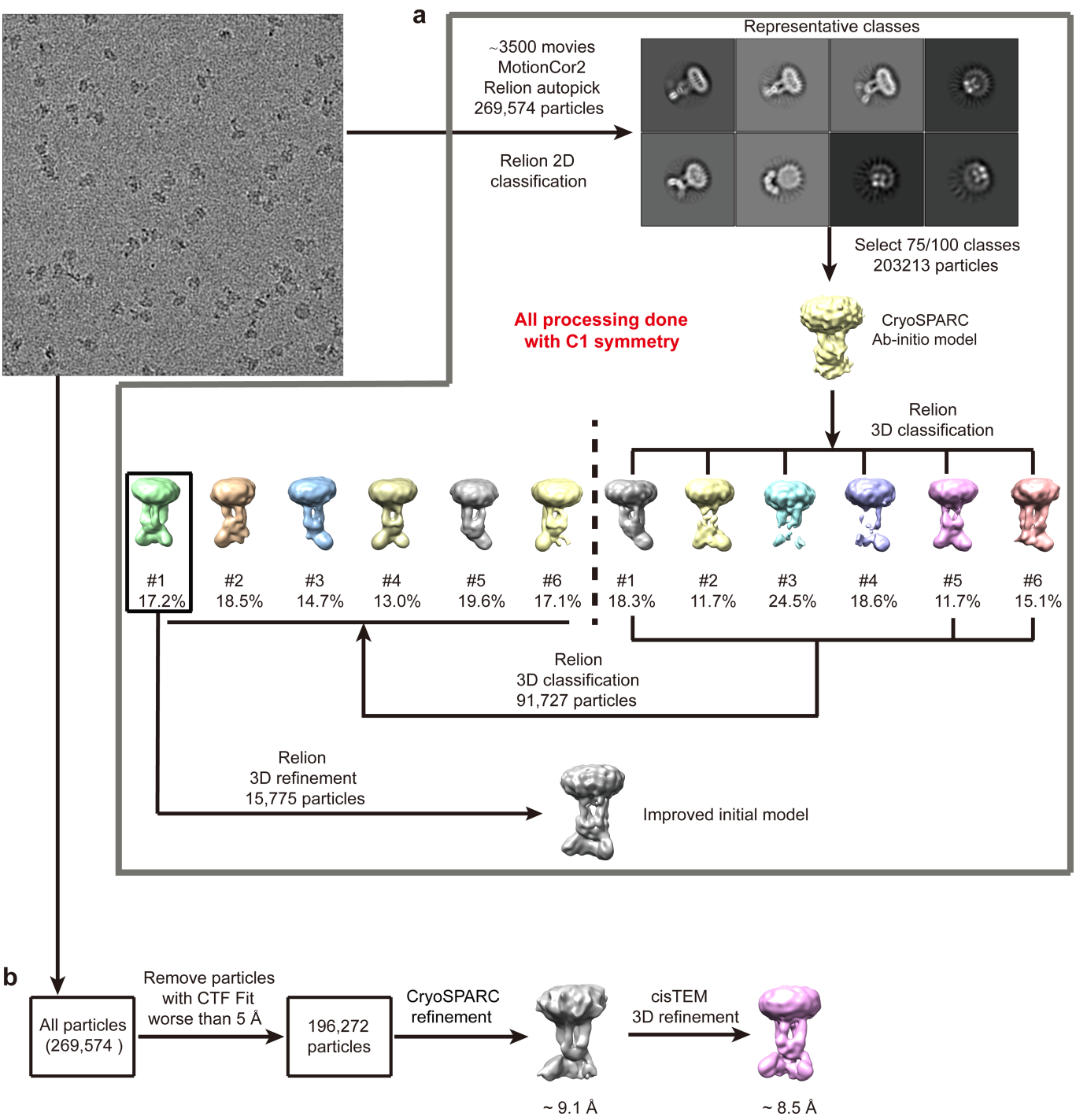
Extended Data Fig. 2 | Functional analysis of MCU. a, **b**, Representative Ca^{2+} uptake experiments using digitonin-permeabilized MCU/EMRE knockout cells without transfection (red line), or expressing zebrafish MCU and EMRE (black line), or expressing zebrafish $MCU_{\Delta NTD}$ and EMRE (grey line). Blue arrows indicate additions of 5 μM $CaCl_2$. A decrease in fluorescence following addition of Ca^{2+} is indicative of Ca^{2+} uptake (for example, black and grey traces). Ruthenium red (RuRed), a proton ionophore (CCCP), or a Ca^{2+} ionophore (ETH129) were added as controls towards the end of each experiment as indicated. RuRed prevents Ca^{2+} uptake, ETH129 demonstrates that the mitochondria are intact, and CCCP demonstrates that Ca^{2+} uptake is dependent upon the proton gradient. Analogous distinct experiments were repeated a total of six times and yielded similar results. **c**, Rho-1D4 western blot analysis of the cell lysates from **a**, demonstrating expression of MCU and EMRE, which were C-terminally tagged with a 1D4 peptide (see Methods). Repeated twice with similar results. **d**, **f**, Representative mitochondrial Ca^{2+} uptake experiments using MCU/EMRE knockout cells for the indicated mutations of human MCU when co-transfected with EMRE (from which the data shown in Fig. 4d are derived, see Methods). Independent experiments were repeated with similar results: $n = 8$ for wild type, $n = 9$ for M263A, and $n = 3$ for the remainder of the mutants. **e**, **g**, Full-size western blots of the cell lysates from **d** and **f** showing protein expression levels (corresponding to Fig. 4d; detected using a Rho-1D4 antibody). Repeated twice with similar results. **h**–**j**, Subunit stoichiometry analysis of MCU proteins using crosslinking. **h**, Crosslinking of purified CeMCU in the detergent *n*-dodecyl-β-D-maltoside (DDM). Indicated concentrations of the crosslinker bis(sulfosuccinimidyl) suberate (BS-3) were incubated with purified CeMCU. Analysis is by SDS–PAGE and coomassie stain. Molecular weight standards are located in the first lane and their positions indicated. The calculated molecular weight of CeMCU is 39.6 kDa based upon its amino acid sequence. Repeated three times with similar results. **i**, Crosslinking of human MCU expressed in HEK293 membranes. Indicated concentrations of the membrane-permeable crosslinker disuccinimidyl glutarate (DSG) were used and human MCU was detected by western blot using a C-terminal Rho-1D4 antibody tag (Methods). Molecular weight standards are located in the first lane and their positions indicated. We note that some oligomerization of human MCU was observed without crosslinker; this phenomenon has been observed previously for human MCU[32]. Repeated four times with similar results. **j**, Crosslinking of purified zebrafish MCU in the detergent digitonin. Indicated concentrations of BS-3 crosslinker were used. Samples were analysed by SDS–PAGE using coomassie stain. Molecular weight standards are located in the last lane and their positions are indicated. Asterisks indicate protein impurities. Repeated twice with similar results.

© 2018 Springer Nature Limited. All rights reserved.



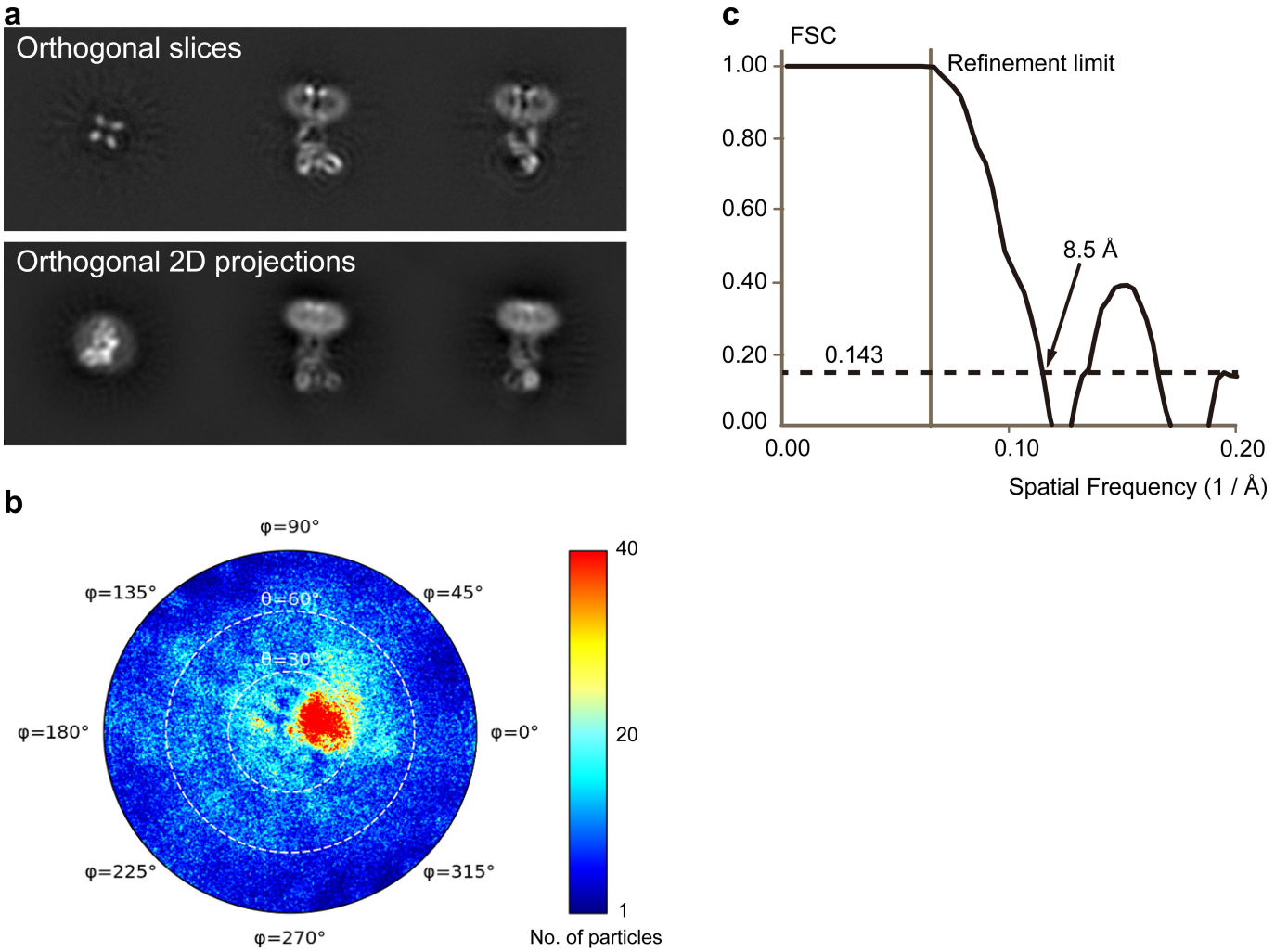
Extended Data Fig. 3 | Flowchart for cryo-EM data processing of CeMCU. Details can be found in the Methods.

© 2018 Springer Nature Limited. All rights reserved.



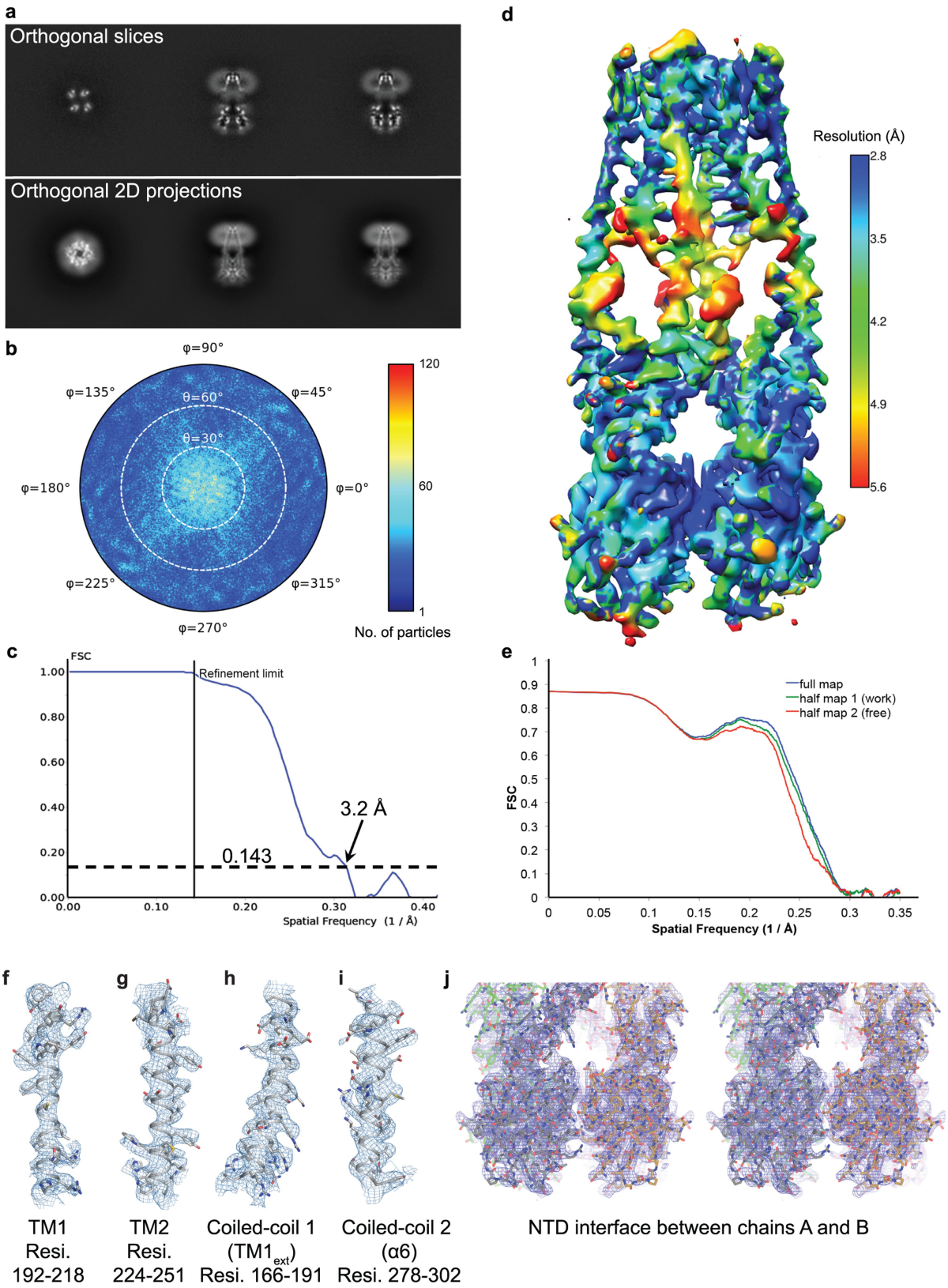
Extended Data Fig. 4 | Flowchart for cryo-EM data processing of zebrafish MCU. a, Initial model generation and improvement. **b**, 3D refinement using the improved initial model from **a**. Details can be found in the Methods.

© 2018 Springer Nature Limited. All rights reserved.



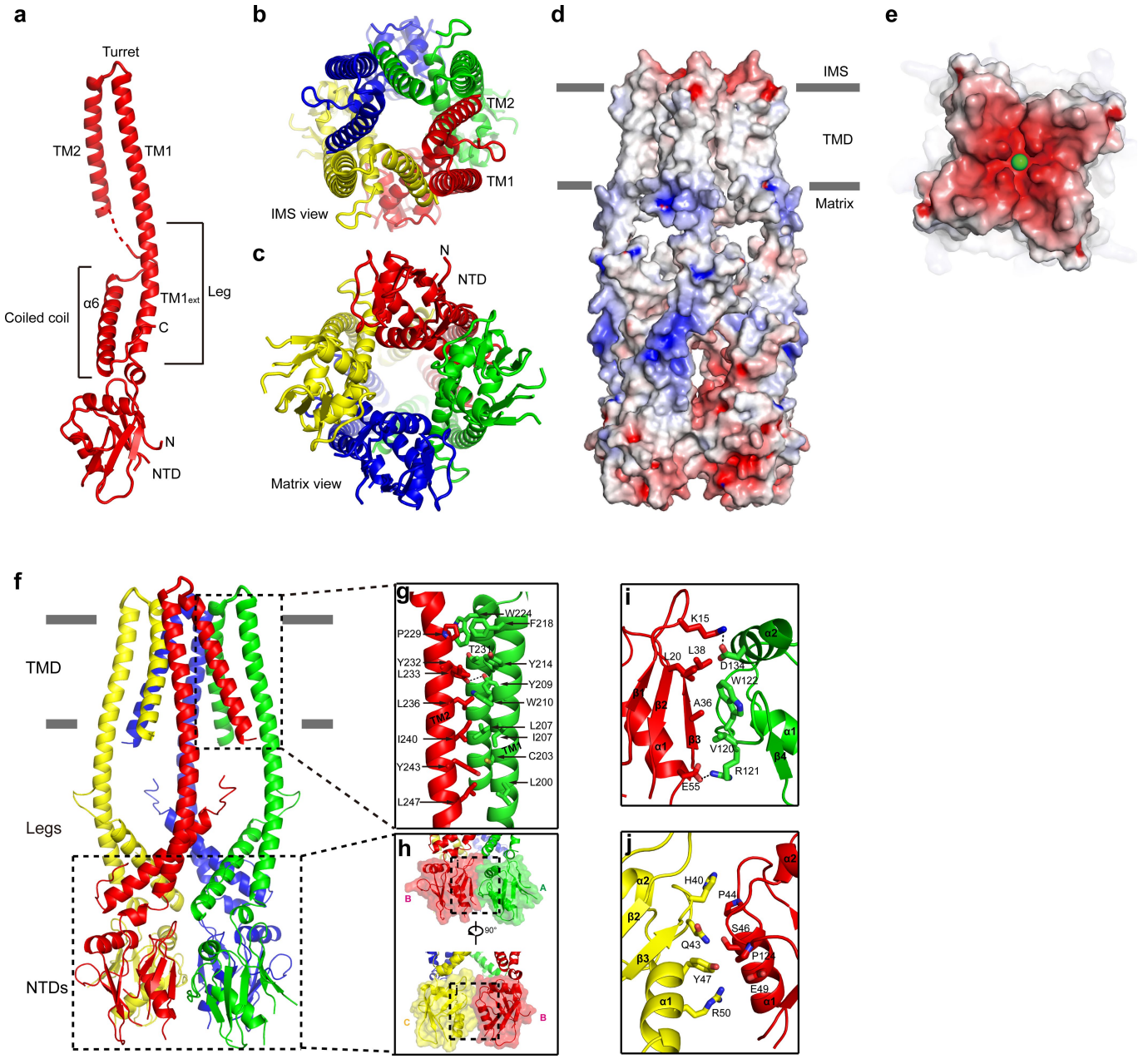
Extended Data Fig. 5 | Cryo-EM reconstruction of zebrafish MCU. **a**, Orthogonal slices (top) and orthogonal 2D projections (bottom) of the final 3D reconstruction (from cisTEM). **b**, Angular orientation distribution of the particles used in the final reconstruction. The particle distribution is indicated by different colour shades. **c**, Gold-standard FSC curve of the final 3D reconstruction. The resolution is ~8.5 Å at the FSC cutoff of 0.143. A thin vertical line indicates that only spatial frequencies to 1/(15 Å) were used to determine particle alignment parameters during refinement in cisTEM.

© 2018 Springer Nature Limited. All rights reserved.



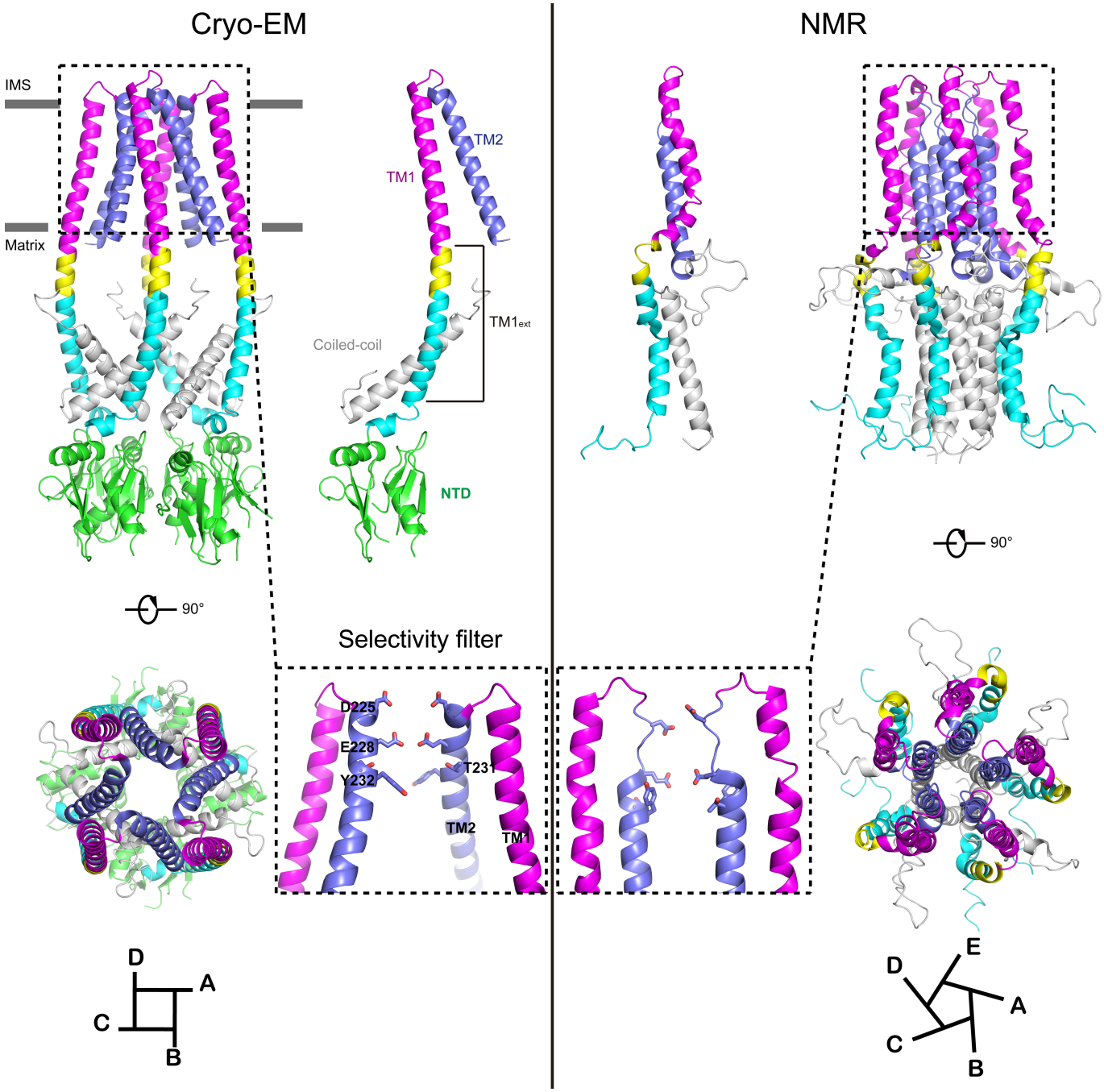
Extended Data Fig. 6 | Cryo-EM structure determination and density of CeMCU. a, Orthogonal slices (top) and orthogonal 2D projections (bottom) of the final 3D reconstruction (from cisTEM). **b**, Angular orientation distribution of the particles used in final reconstruction. The particle distribution is indicated by different colour shades. **c**, Gold-standard FSC curve of the final 3D reconstruction. The resolution is 3.2 Å at the FSC cutoff of 0.143. A thin vertical line indicates that only spatial frequencies to 1/(7 Å) were used to determine particle alignment parameters during refinement. **d**, Local resolution of the map estimated using the blocres program and coloured as indicated. **e**, Model validation. Comparison of the FSC curves between model and half map 1 (work), model and half map 2 (free), and model and full map are plotted in green, red and blue, respectively. **f–i**, Densities (8σ contours) of α-helical regions of CeMCU are shown in the context of the atomic model with side chains shown as sticks and the backbone as ribbons. **j**, Stereo representation of the density (8σ contour) of the NTDs, with the interface between chains A and B in the foreground.

© 2018 Springer Nature Limited. All rights reserved.



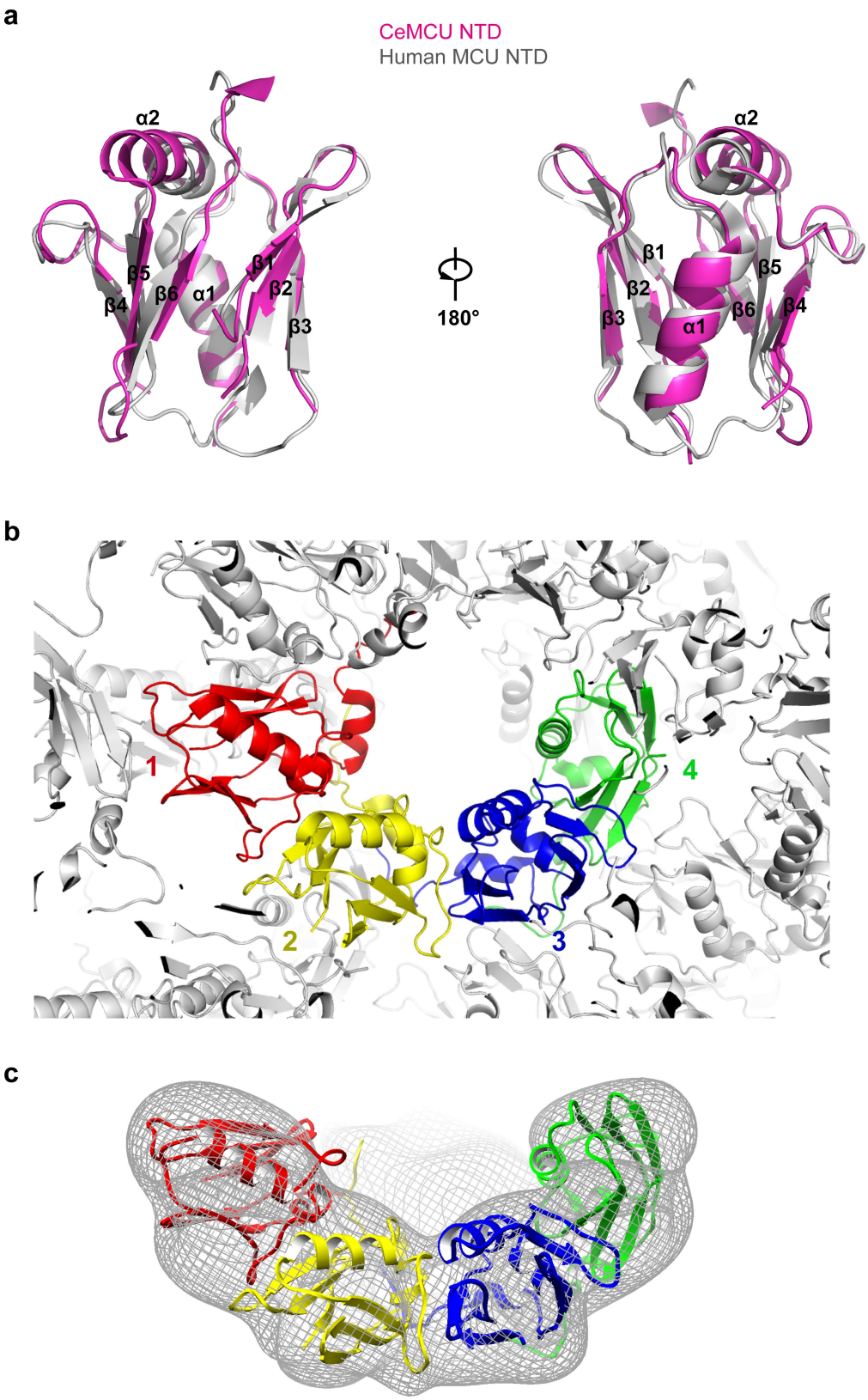
Extended Data Fig. 7 | Structure of CeMCU. a, Ribbon representation of an isolated subunit from Fig. 1c, viewed parallel to the membrane, with the structural features labelled. Disordered regions connecting TM2 and α6 are indicated as dashed lines. **b**, **c**, IMS and matrix views of Fig. 1c. **d**, **e**, Molecular surface coloured according to electrostatic potential (red, -8 kT e^{-1}; white, neutral; blue, $+8$ kT e^{-1}). **d**, Shown in the same orientation as Fig. 1c with approximate membrane boundaries as grey bars. **e**, IMS view. A green sphere indicates the position of Ca^{2+} in site 1. **f**, Overall structure of CeMCU, depicted similarly to Fig. 1. **g**, Subunit interactions within the TMD. Ribbon representations of TM1 and TM2 from one subunit (green) and TM2 of the neighbouring subunit (red) are shown. Residues participating in van der Waals or hydrogen bonding interactions are shown as sticks; dashed lines indicate hydrogen bonds. Atom colouring: nitrogen, blue; oxygen, red; and sulfur, green. **h**, Interfaces within the NTDs. The NTDs of subunit A, B and C (green, red and yellow, respectively) are shown in ribbon and surface representations. Each NTD contains six β-strands (β1 through β6) and two α-helices (α1 and α2). There are four interfaces between the NTDs, and these are of two types. Interface 1 (for example, between protomers A and B or C and D in the atomic model) consists of an interaction between β3 of one NTD and the β1–β2 loop of another; this interface buries 2,188 $Å^2$ of molecular surface in the assembled channel. Interface 2 (for example, between protomers A and D or B and C) is less extensive, burying 1,440 $Å^2$ total surface area, and involves contacts located near the N-terminal ends of the α1 helices from two NTDs. **i**, **j**, Details of interfaces 1 and 2 between NTDs, respectively.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 8 | Comparison with an NMR structure (PDB: 5ID3). Left, various representations of the cryo-EM structure of CeMCU. The NTD, $TM1_{ext}$, TM1, TM2 and coiled-coil regions are shown in different colours. Right, structure deduced from NMR studies of *C. elegans* MCU-ΔNTD (PDB: 5ID3). It is coloured according to the cryo-EM structure using the sequence alignment shown in Extended Data Fig. 1. The boxed regions highlight differences in vicinity of the WDXXEP signature sequence and pore. Pore-lining residues (D225, E228, T231 and Y232; *C. europaea* numbering) are shown as sticks.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | Structural comparison of the NTDs of CeMCU, zebrafish MCU and human MCU. a, An NTD from the cryo-EM structure of CeMCU (magenta) is superimposed with the crystal structure of an isolated NTD from human MCU (grey, PDB: 5KUG, r.m.s.d. = 2.0 Å). The secondary structure features are indicated; two views are shown. **b**, Crystal lattice from an X-ray structure of an isolated NTD of human MCU (PDB: 4XTB). Four neighbouring NTDs are coloured and other NTDs are grey. A similar arrangement is present in PDB 5KUG. **c**, Matrix view of the cryo-EM reconstruction of zebrafish MCU (from Fig. 2d).

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	CeMCU (EMD-7971) (PDB 6DNF)	Zebrafish MCU (EMDB-7972)
Data collection and processing		
Microscope	FEI Titan Krios (at MSKCC)	FEI Titan Krios (at MSKCC)
Camera	Gatan K2 Summit	Gatan K2 Summit
Magnification	29,000×	37,000×
Voltage (kV)	300	300
Electron exposure (e^{-}/Å^2)	80	100.2
Defocus range (μm)	−0.9 ~ −2.9	−1.2 ~ −3.2
Pixel size (Å)	0.8613 (0.4307)	0.692 (0.346)
Software	RELION 2.1, cisTEM 1.0	RELION 2.1, cryoSPARC, cisTEM 1.0,
Symmetry imposed	C2	C1
Initial particle images (no.)	547,637	269,574
Final particle images (no.)	376,541	196,272
Overall map resolution (Å)	3.2	8.5
FSC threshold 0.143		
Local map resolution range (Å)	2.8-5.6	
Refinement		
Software	Phenix 1.13 real-space-refine	
Initial model used (PDB code)	N/A	
Model resolution (Å)	4.0	
FSC threshold 0.5		
Map sharpening *B* factor (Å^2)	−90	−90
Model composition		
Non-hydrogen atoms	7,312	
Protein residues	900	
Ligands	3	
Lipids	8	
B factors (Å^2)		
Protein	136.7	
Ligand	153.9	
R.m.s. deviations		
Bond lengths (Å)	0.016	
Bond angles (°)	0.830	
Validation		
MolProbity score	1.26	
Clashscore	1.82	
Poor rotamers (%)	0.0	
Ramachandran plot		
Favored (%)	95.4	
Allowed (%)	4.6	
Disallowed (%)	0.0	

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0329-2

The formation of solar-neighbourhood stars in two generations separated by 5 billion years

Masafumi Noguchi[1]*

The chemical compositions of stars encode those of the gas from which they formed, providing important clues regarding the formation histories of galaxies. A powerful diagnostic is the abundance of α elements (O, Mg, Si, S, Ca and Ti) relative to iron, [α/Fe]. The α elements are synthesized and injected into the interstellar medium by type II supernovae, which occur about ten million years after their originating stars form; by contrast, iron is returned to the interstellar medium by type Ia supernovae, which occur after a much longer timescale of roughly one billion years[1]. Periods of rapid star formation therefore tend to produce high-[α/Fe] stellar populations (because only type II supernovae have time to contribute to interstellar-medium enrichment as the stellar population forms), whereas low-[α/Fe] stars require periods of star formation that last more than a few billion years (over which timescales type Ia supernovae begin to affect the elemental composition of the interstellar medium more strongly than type II supernovae). The existence of two distinct groups of stars in the solar neighbourhood[2–7], one with high [α/Fe] and the other with low [α/Fe], therefore suggests two different origins, but the mechanism by which this bimodal distribution arose remains unknown. Here we use a model of disk-galaxy evolution to show that the two episodes of star formation[8] predicted by the 'cold flow' theory of galactic gas accretion[9,10] also explain the observed chemical bimodality. In this scenario, the high-[α/Fe] stars form early, during an initial phase of accretion that involves infalling streams of cold primordial gas. There is then a hiatus of around two billion years until the shock-heated gas in the galactic dark-matter halo has cooled as a result of radiation and can itself commence accretion. The low-[α/Fe] stars form during this second phase. The peaks in these two star-formation episodes are separated by around five billion years. In addition, the large-scale variation in the abundance patterns of these two stellar populations that has been observed for the Milky Way[5,7] is partially explained by the spatial variation in this gas-accretion history.

The Milky Way galaxy provides an excellent test bed for models of the chemical evolution of galaxies because its individual stars can be resolved and examined in detail, unlike most other galaxies. The bimodality[2–7] that has been observed in the chemical composition of stars in the solar neighbourhood is generally assumed to suggest (although some scepticism remains[11]) that star formation in this region did not proceed continuously but halted for a certain period, leading to two separated episodes of star formation[12,13]. How star formation proceeds is largely governed by how the primordial gas accretes onto the forming galaxy. The longstanding paradigm—the shock-heating theory—argues that the gas that entered the dark-matter halo is heated to a high temperature by a shock wave that provides sufficient pressure for the gas to maintain equilibrium within the gravitational field of the halo. Then, the gas cools by emitting radiation and accretes gradually to the forming galaxy at the centre (hot-mode accretion)[14]. On the basis of detailed numerical simulations, a different hypothesis has been put forward that proposes that a substantial part of the primordial gas remains cold and reaches the galaxy in narrow streams in an almost freefall fashion (the 'cold flow' theory)[9,10]. A hydrodynamical simulation in growing dark-matter haloes[8] predicts the presence of a gap in gas accretion (figures 2–4 of ref. [8]) and the resulting bimodal history of

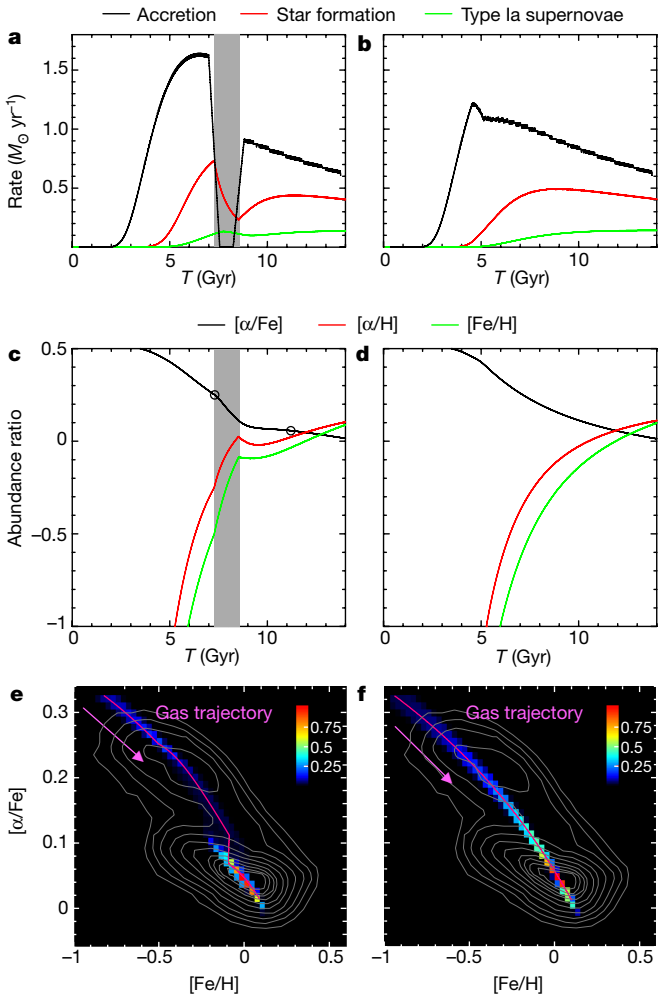


Fig. 1 | Chemical evolution and the chemical abundance diagram for the solar neighbourhood. a, **c**, **e**, Cold-flow model. **b**, **d**, **f**, Shock-heating model. All panels refer to the region 7 kpc < R < 9 kpc, where R is the distance from the Galactic centre. T is the time since the beginning of the Universe. The accretion and star formation rates are expressed in masses of the Sun per year. The type Ia supernova rate is the number per century. The grey shaded region in **a** and **c** highlights the gap in gas accretion. Open circles in **c** indicate [α/Fe] at the two star-formation peaks. **e**, **f**, Stellar distribution in the chemical diagram, with the colour scale indicating the relative frequency with respect to the maximum value in each panel. The contours are observed distributions based on the APOGEE data[4]. The purple line indicates the evolution of the gas, with the arrow indicating the time sequence.

[1]Astronomical Institute, Tohoku University, Sendai, Japan. e-mail: noguchi@astr.tohoku.ac.jp

© 2018 Springer Nature Limited. All rights reserved.

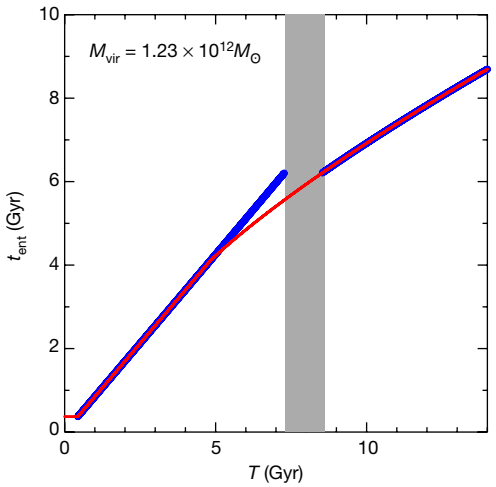


Fig. 2 | Time of entering the dark-matter halo as a function of arrival time to the disk. For gas that arrives at the disk at time T, t_{ent} denotes the time at which it enters the dark-matter halo. Blue and red lines indicate the cold-flow and shock-heating models, respectively. The grey shaded region corresponds to the gap in gas accretion in Fig. 1a, c. $M_{vir} = 1.23 \times 10^{12} M_\odot$ is the current virial mass used in the models, where $M_\odot$ is the mass of the Sun.

disk growth (figures 9–11 of ref. [8]), providing a viable basis for a more detailed understanding of chemical evolution.

This work, based on these findings, takes this idea one step further by incorporating the star-formation process and chemical evolution in a simple model of disk-galaxy evolution, with the aim of clarifying the origin of the observed chemical bimodality. In the model, a growing dark-matter halo gathers the surrounding primordial gas, which subsequently accretes to the disk component of the galaxy and forms stars[15]. We construct two models with different gas-accretion histories, corresponding to the shock-heating and cold-flow scenarios.

Figure 1 depicts the time evolution of the two models and the resulting [α/Fe]–[Fe/H] diagram (referred to as the chemical diagram hereafter) for the solar neighbourhood. The cold-flow theory predicts a clear chemical bimodality, whereas the shock-heating model produces a single sequence (Fig. 1e, f). The locations of the two stellar groups in the cold-flow model agree with those observed by APOGEE[4]. The period of dormant accretion about 6 Gyr ago (shaded part in Fig. 1a) is an essential event that contributes to creating the well-defined bimodality in the cold-flow model. How this gap arises is clarified in Fig. 2. We define t_{ent} to be the time when the primordial gas enters the dark-matter halo and T as the cosmological time from the beginning of the Universe. At $t_{ent} \approx 4.3$ Gyr, the shock-heating model starts to accrete gas on the timescale of cooling, as opposed to that of freefall, and its accretion is delayed with respect to the freefall accretion in the cold-flow model. In the cold-flow model, the gas that arrives at the disk at $T \approx 8.6$ Gyr in the shock-heating model is already accreting at $T \approx 7.3$ Gyr. At $t_{ent} \approx 6.3$ Gyr, the cold mode of accretion ends and the gas accretion is switched to the hot mode in the cold-flow model. Therefore, the gas that enters the halo after $t_{ent} \approx 6.3$ Gyr will not accrete until $T \approx 8.6$ Gyr. This change in the accretion mode gives rise to a period in which accretion is halted (shaded region in Fig. 2), as documented in the previous numerical study[8]. After $t_{ent} \approx 6.3$ Gyr, the two models accrete gas on the timescale of cooling. Although the detailed times have a significant uncertainty of 1–2 Gyr depending on the assumed galactic mass, the prediction of the gap in the cold-flow model is robust.

Figure 1 details the evolution as dictated by the cold-flow model. The first phase of gas accretion induces a brief period of active star formation with a half peak width of about 2 Gyr, and α elements start to be released immediately by type II supernovae. This star formation begins to weaken before a large amount of iron starts to be supplied by type Ia supernovae. Therefore, most stars formed during this epoch have high [α/Fe]. During the subsequent period in which star formation weakens, type Ia supernovae start to dominate. Because the enrichment of α elements is overtaken by iron enrichment, [α/Fe] in the interstellar medium decreases further. When the second phase of gas accretion starts about 2 Gyr later and star formation is activated again, [α/Fe] stops changing rapidly and decreases very slowly thereafter. The second-generation stars have lower [α/Fe] by about 0.2 dex than the first-generation stars (two open circles in Fig. 1c). The paucity of stars with intermediate [α/Fe] creates a gap between the high- and low-[α/Fe] sequences in the chemical diagram. The star-formation history in the solar neighbourhood deduced empirically from the observed chemical diagram using a galactic chemical evolution code[13] reveals a gap about 7 Gyr ago, in agreement with the prediction of the cold-flow model.

Recent large-scale surveys such as GAIA-ESO[16], APOGEE[17] and HARPS GTO planet search[18] extend the stellar-abundance study to a sizable fraction of the Milky Way, reaching more than 10 kpc from the Sun. The prediction of the cold-flow model can be compared to the observations for this large-scale region of the Galactic disk. Figure 3g–i shows the distribution of stars in the chemical diagram for three different regions: the inner disk, the solar neighbourhood and the outer disk.

In view of the idealized nature of the model, here we focus on the peak positions and the relative dominance of the two stellar groups. The low-[α/Fe] sequence increases its dominance over the high-[α/Fe] one with increasing distance from the Galactic centre, in agreement with obseravtions[5,7]. This is because the second phase of accretion, which increasingly dominates (Fig. 3a–c), involves the gas with higher angular momentum, which accretes onto larger radii. This behaviour is consistent with the result from a more detailed simulation[19]. We also note that the [α/Fe] ratio of the low-[α/Fe] sequence increases with radius in the model, whereas its [Fe/H] decreases. This is because the outer disk is still in a younger phase of chemical evolution at present, making stars from interstellar medium that is not yet heavily polluted by iron (Fig. 3f). The decrease in [Fe/H] with radius from 0.13 to 0.01 to −0.26 is in fair agreement with the observed decrease from 0.2 to 0.05 to −0.3, although the model [α/Fe] varies more strongly than the observation. This may be due in part to the idealization of the model, which assumes that the stellar yields and star-formation efficiency are constant, whereas there is evidence that these quantities can be variable[20,21]. The model is not intended to be complete at this stage.

The two stellar groups are closely related to different structural components of the Milky Way[7]. Most high-[α/Fe] stars belong to the thick-disk component. The low-[α/Fe] stars tend to have small velocities perpendicular to the plane of the disk and hence contribute to the thin disk. There is an argument[22,23] that the formation of thick and thin disks is not strictly sequential but partly overlapped in time. If this is the case, then the simple time sequence of chemical evolution shown in the model should be revised. More realistic but expensive simulations that treat chemo-dynamical evolution directly[24] are required to address these kinematic issues.

Recent advances in observational techniques make the galaxies in the Local Group and other nearby galaxies feasible targets for studying star-formation history by examining the colour–magnitude diagram of their member stars[25–27]. Because these spiral galaxies span a wide range of masses and the cold-flow theory predicts different evolutions depending on galactic mass, this nearby sample of galaxies provides powerful diagnostics for the theory.

M31, the largest galaxy in the Local Group, seems to have experienced two episodes of star formation separated by a break about 6 Gyr ago[25]. The cold-flow theory predicts that the quiescent period between the two star-formation phases becomes longer for more massive galaxies; the break observed in M31 seems to be more prominent than the one inferred for the Milky Way[25], in agreement with this prediction. M33 and NGC 300, which are several times less massive than the Milky Way, have signatures of inside-out formation and their star-formation histories seem to be monotonic[26,27]. According to the cold-flow model, the Milky Way has approximately the minimum mass for which a clear star-formation gap is expected. Observations of M33 and NGC 300[26,27] appear to support this prediction, although the data include a large uncertainty.

© 2018 Springer Nature Limited. All rights reserved.

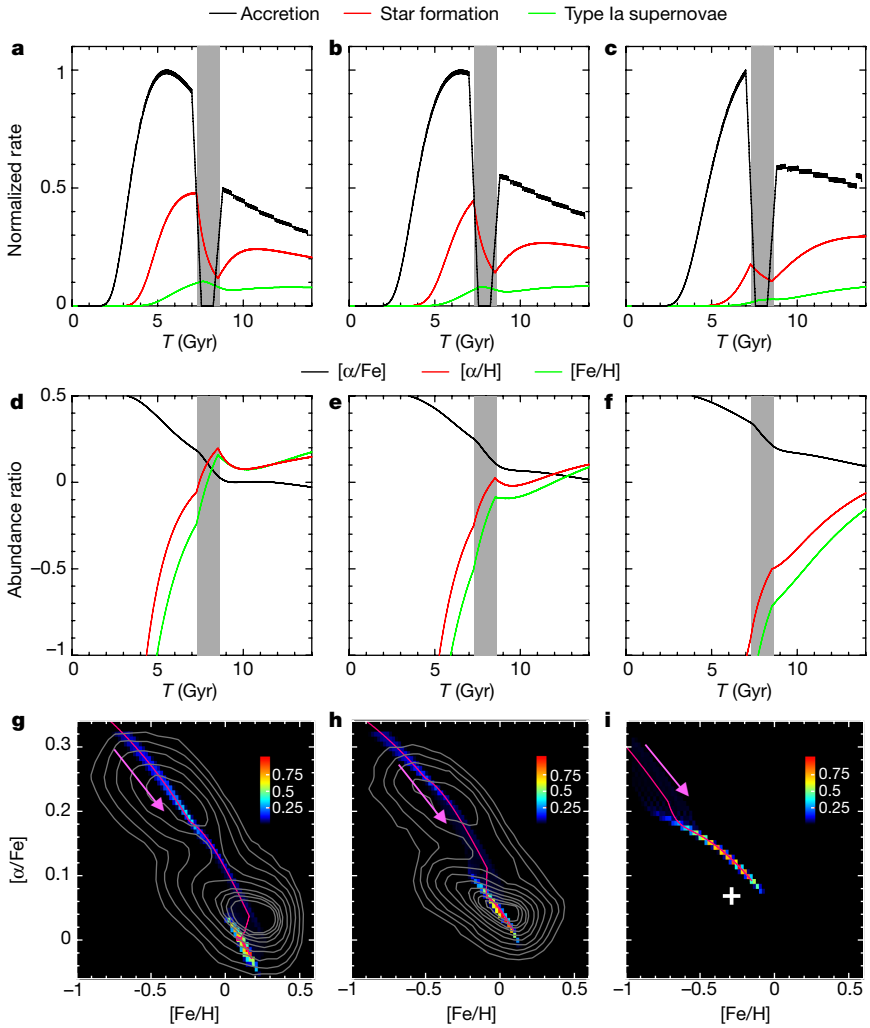


Fig. 3 | Evolution of the cold-flow model in three different zones. a–i, Same as Fig. 1, but for three regions of the disk: 5 kpc < R < 7 kpc (**a**, **d**, **g**), 7 kpc < R < 9 kpc (**b**, **e**, **h**; identical to Fig. 1a, c, e) and 11 kpc < R < 13 kpc (**c**, **f**, **i**). The rates in **a**–**c** are normalized so that the maximum value of the accretion rate in each panel is equal to 1. The contours in **g** and **h** are based on the APOGEE data[4]. The cross in **i** indicates the approximate location of the low-[α/Fe] sequence inferred from APOGEE data[5,7] for the outer disk. Only one [α/Fe] sequence is discernible in the model in **i**.

Detailed simulations suggest that the evolution of disk galaxies is a complex process that involves the interplay of multiple pathways[28]. For example, the formation of the high-[α/Fe] sequence by gas-rich mergers at early times is recognized in some cases[28]. Radial migration of high-[α/Fe] stars may contribute to bimodality[11]. The simple model used here isolates the effect of cold flows and highlights their role. The cold-flow hypothesis can explain the red colour of massive elliptical galaxies, in combination with the energy feedback from active galactic nuclei[29]. It also provides a plausible interpretation for the high incidence of strongly star-forming galaxies at high redshifts[30]. Finally, the cold-flow model not only is applicable to a large population of distant galaxies but also provides a possible explanation for the diversity in nearby individual galaxies, including the Milky Way.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0329-2.

Received: 30 January 2018; Accepted: 11 June 2018;
Published online 25 July 2018.

1. Matteucci, F. & Greggio, L. Relative roles of type I and II supernovae in the chemical enrichment of the interstellar gas. *Astron. Astrophys.* **154**, 279–287 (1986).
2. Gratton, R., Carretta, E., Matteucci, F. & Sneden, C. The [Fe/O] ratio in field stars and the history of star formation of the solar neighbourhood. *ASP Conf. Ser.* **92**, 307–316 (1996).
3. Adibekyan, V. Zh. et al. Chemical abundances of 1111 FGK stars from the HARPS GTO planet search program: galactic stellar populations and planets. *Astron. Astrophys.* **545**, A32 (2012).
4. Haywood, M. et al. When the Milky Way turned off the lights: APOGEE provides evidence of star formation quenching in our Galaxy. *Astron. Astrophys.* **589**, A66 (2016).
5. Anders, F. et al. Chemodynamics of the Milky Way I. The first year of APOGEE data. *Astron. Astrophys.* **564**, A115 (2014).
6. Bensby, T., Feltzing, S. & Oey, M. S. Exploring the Milky Way stellar disk: a detailed elemental abundance study of 714 F and G dwarf stars in the solar neighbourhood. *Astron. Astrophys.* **562**, A71 (2014).
7. Hayden, M. R. et al. Chemical cartography with APOGEE: metallicity distribution functions and the chemical structure of the Milky Way disk. *Astrophys. J.* **808**, 132 (2015).
8. Birnboim, Y., Dekel, A. & Neistein, E. Bursting and quenching in massive galaxies without major mergers or AGNs. *Mon. Not. R. Astron. Soc.* **380**, 339–352 (2007).
9. Fardal, M. A. et al. Cooling radiation and the Lyα luminosity of forming galaxies. *Astrophys. J.* **562**, 605–617 (2001).
10. Dekel, A. & Birnboim, Y. Galaxy bimodality due to cold flows and shock heating. *Mon. Not. R. Astron. Soc.* **368**, 2–20 (2006).
11. Schönrich, R. & Binney, J. Chemical evolution with radial mixing. *Mon. Not. R. Astron. Soc.* **396**, 203–222 (2009).
12. Chiappini, C., Matteucci, F. & Gratton, R. The chemical evolution of the Galaxy: the two-infall model. *Astrophys. J.* **477**, 765–780 (1997).
13. Snaith, O. et al. Reconstructing the star formation history of the Milky Way disc(s) from chemical abundances. *Astron. Astrophys.* **578**, A87 (2015).
14. Rees, M. J. & Ostriker, J. P. Cooling, dynamics and fragmentation of massive gas clouds: clues to the masses and radii of galaxies and clusters. *Mon. Not. R. Astron. Soc.* **179**, 541–559 (1977).

© 2018 Springer Nature Limited. All rights reserved.

15. Noguchi, M. Possible imprints of cold mode accretion on the present-day properties of disk galaxies. *Astrophys. J.* **853**, 67 (2018).
16. Gilmore, G. et al. The Gaia-ESO public spectroscopic survey. *Messenger* **147**, 25–31 (2012).
17. Majewski, S. R. et al. The Apache Point Observatory Galactic Evolution Experiment (APOGEE). *Astron. J.* **154**, 94–139 (2017).
18. Mayor, M. et al. Setting new standards with HARPS. *Messenger* **114**, 20–24 (2003).
19. Danovich, M., Dekel, A., Hahn, O., Ceverino, D. & Primack, J. Four phases of angular-momentum buildup in high-*z* galaxies: from cosmic-web streams through an extended ring to disc and bulge. *Mon. Not. R. Astron. Soc.* **449**, 2087–2111 (2015).
20. Romano, D., Karakas, A. I., Tosi, M. & Matteucci, F. Quantifying the uncertainties of chemical evolution studies II. Stellar yields. *Astron. Astrophys.* **522**, A32 (2010).
21. Wyse, R. F. G. & Silk, J. Star formation rates and abundance gradients in disk galaxies. *Astrophys. J.* **339**, 700–711 (1989).
22. Rojas-Arriagada, A. et al. The Gaia-ESO Survey: exploring the complex nature and origins of the Galactic bulge populations. *Astron. Astrophys.* **601**, A140 (2017).
23. Haywood, M., Di Matteo, P., Lehnert, M. D., Katz, D. & Gómez, A. The age structure of stellar populations in the solar vicinity: clues of a two-phase formation history of the Milky Way disk. *Astron. Astrophys.* **560**, A109 (2013).
24. Kobayashi, C. & Nakasato, N. Chemodynamical simulations of the Milky Way galaxy. *Astrophys. J.* **729**, 16 (2011).
25. Williams, B. F. et al. PHAT XIX. The ancient star formation history of the M31 disk. *Astrophys. J.* **846**, 145 (2017).
26. Williams, B. F., Dalcanton, J. J., Dolphin, A. E., Holtzman, J. & Sarajedini, A. The detection of inside-out disk growth in M33. *Astrophys. J.* **695**, L15–L19 (2009).
27. Gogarten, S. M. et al. The Advanced Camera for Surveys Nearby Galaxy Survey Treasury. V. Radial star formation history of NGC 300. *Astrophys. J.* **712**, 858–874 (2010).
28. Grand, R. J. J. et al. Origin of chemically distinct discs in the Auriga cosmological simulations. *Mon. Not. R. Astron. Soc.* **474**, 3629–3639 (2018).
29. Cattaneo, A., Dekel, A., Devriendt, J., Guiderdoni, B. & Blaizot, J. Modelling the galaxy bimodality: shutdown above a critical halo mass. *Mon. Not. R. Astron. Soc.* **370**, 1651–1665 (2006).
30. Dekel, A. et al. Cold streams in early massive hot haloes as the main mode of galaxy formation. *Nature* **457**, 451–454 (2009).

Acknowledgements We acknowledge I. Ferreras for comments.

Reviewer information *Nature* thanks G. Cescutti, A. Dekel and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Competing interests The author declares no competing interests.

Additional information
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.N.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

The disk-galaxy evolution model incorporating type Ia supernovae. The Milky Way model used in this analysis was obtained by using a code that divides a disk galaxy into a series of concentric annuli and calculates the time evolution of masses of gas and stars in each annulus under the specified gas-accretion history and star-formation law[15]. Following the result of cosmological numerical simulations[31], the growth of the virial mass of the dark-matter halo with redshift z is specified as $M_{\rm vir}(z) = M_{\rm vir,0}{\rm e}^{-2Kz/C}$, where $M_{\rm vir,0}$ is the present virial mass, $K = 3.7$ is a constant[32] and the halo concentration parameter C at the present epoch is given by

$$\log({\rm C}) = 0.971 - 0.094\log[M_{\rm vir,0}/(10^{12}h^{-1}M_{\odot})]$$

where $h = H_0/(100\,{\rm km\,s^{-1}\,Mpc^{-1}})$ and H_0 is the Hubble constant. The primordial gas gathered by the dark-matter halo is assumed to have the same density profile as the halo, described by the Navarro–Frenk–White profile.

The shock-heating model is constructed by assuming that the gas is heated to the virial temperature when entering the host halo and accretes to the disk plane with the timescale of radiative cooling (or with the freefall time if the cooling time is shorter). The cold-flow model is based on the Dekel–Birnboim model[10]. On the basis of stability analysis and aided by numerical simulations, it provides information on when a stable shock wave appears in the growing halo and the gas accretion changes from cold mode to hot mode. The gas is made to accrete with the freefall time in the initial cold-flow phase, but with the longer of the cooling time and the freefall time after the shock develops. We used the prediction of the original Dekel–Birnboim model[10] for the cold-flow model presented here instead of the extended cold-accretion model[15]. The latter model is a hypothetical model devised to match the observed large disk sizes of galaxies much more massive than the Milky Way (stellar mass larger than $10^{11}M_{\odot}$), for which the original Dekel–Birnboim model predicts too small radii. On the other hand, the Dekel–Birnboim model is based on a detailed stability analysis for virial shock and is confirmed by numerical simulation for collapsing gas spheres with radiative cooling. The Milky Way and other nearby spiral galaxies lie in the mass range for which the Dekel–Birnboim model is applicable.

The original code considers the chemical enrichment of the interstellar gas attributed only to type II supernovae. Here, chemical enrichment due to type Ia supernovae is added. Specifically, α elements of $3.81M_{\odot}$ and iron of $0.094M_{\odot}$ are added to the cold gas for one type II supernova. One type Ia supernova is assumed to return $1.65M_{\odot}$ of material to the cold gas, of which α elements and iron make up $0.438M_{\odot}$ and $0.88M_{\odot}$, respectively. The yields for α elements by type II supernovae and for iron by type Ia supernovae adopted here are both increased by 20% from the values listed in the literature[33] to obtain the best match to the observation. Type II supernovae are assumed to explode instantaneously after the birth of progenitor stars. Type Ia supernovae take a certain period after stellar birth before they explode. An observationally estimated delayed time distribution (DTD)[34] is introduced so that the rate for type Ia supernovae (SN) is

$${\rm DTD}(t) = 1\times10^{-3}\frac{\rm SN}{M_{\odot}{\rm Gyr}}\left(\frac{t}{\rm Gyr}\right)^{-1.1}$$

where t denotes the time since the progenitor formed. Considering the uncertainty in binary stellar evolution[35], we assume that DTD = 0 for $t < 0.3$ Gyr. The type II supernova feedback efficiency is taken to be $\varepsilon = 0.15$ instead of the original value 0.05, to match the model with the Milky Way; this change does not influence the main conclusion of this work. All other model parameters are the same as those adopted previously[15].

Adopting the current virial mass of the dark-matter halo[36] of $M_{\rm vir,0} = 1.2\times10^{12}M_{\odot}$ (with a corresponding collapse redshift of $2K/C = 0.78$), the cold-flow model produces a stellar disk of mass $6\times10^{10}M_{\odot}$ with a scale length of 3.0 kpc in the present epoch, in broad agreement with the estimate for the Milky Way[37,38]. The inner disk, solar neighbourhood and outer disk analysed here are defined as an annulus in the disk with $5\,{\rm kpc} < R < 7\,{\rm kpc}$, $7\,{\rm kpc} < R < 9\,{\rm kpc}$ and $11\,{\rm kpc} < R < 13\,{\rm kpc}$, respectively, where R is the galactocentric radius. Each region contains 15, 11 and 8 radial bins, leading to the spread of the stellar distribution shown in Figs. 1 and 3. The abundance ratio is the value relative to the solar value on a logarithmic scale. Solar abundances of $[\alpha/{\rm H}] = 2.22\times10^{-2}$ and $[{\rm Fe/H}] = 1.77\times10^{-3}$ are assumed[39]. The contours in Figs. 1 and 3 are taken from ref. [4]; credit: M. Haywood et al. A&A, 589, 66 (2016), reproduced with permission © ESO.

Data availability. The data that support the findings of this study are available from the corresponding author on reasonable request.

Code availability. We have opted not to make available the code used to calculate the evolution of the disk-galaxy models because it is a part of an integrated program currently in use for other projects.

31. Wechsler, R. H., Bullock, J. S., Primack, J. R., Kravtsov, A. V. & Dekel, A. Concentrations of dark halos from their assembly histories. *Astrophys. J.* **568**, 52–70 (2002).
32. Bullock, J. S. et al. Profiles of dark haloes: evolution, scatter and environment. *Mon. Not. R. Astron. Soc.* **321**, 559–575 (2001).
33. Ferreras, I. & Silk, J. Type Ia supernovae and the formation history of early-type galaxies. *Mon. Not. R. Astron. Soc.* **336**, 1181–1187 (2002).
34. Maoz, D., Sharon, K. & Gal-Yam, A. The supernova delay time distribution in galaxy clusters and implications for type-Ia progenitors and metal enrichment. *Astrophys. J.* **722**, 1879–1894 (2010).
35. Claeys, J. S. W., Pols, O. R., Izzard, R. G., Vink, J. & Verbunt, F. W. M. Theoretical uncertainties of the type Ia supernova rate. *Astron. Astrophys.* **563**, A83 (2014).
36. Smith, M. C. et al. The RAVE survey: constraining the local Galactic escape speed. *Mon. Not. R. Astron. Soc.* **379**, 755–772 (2007).
37. Bovy, J. & Rix, H.-W. A direct dynamical measurement of the Milky Way's disk surface density profile, disk scale length, and dark matter profile at $4\,{\rm kpc} < r < 9\,{\rm kpc}$. *Astrophys. J.* **779**, 115 (2013).
38. Licquia, T. C. & Newman, J. A. Improved estimates of the Milky Way's stellar mass and star formation rate from hierarchical Bayesian meta-analysis. *Astrophys. J.* **806**, 96 (2015).
39. Asplund, M., Grevesse, N., Sauval, A. J. & Scott, P. The chemical composition of the Sun. *Annu. Rev. Astron. Astrophys.* **47**, 481–522 (2009).

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0348-z

Spontaneous emission of matter waves from a tunable open quantum system

Ludwig Krinner[1], Michael Stewart[1], Arturo Pazmiño[1], Joonhyuk Kwon[1] & Dominik Schneble[1]*

The decay of an excited atom undergoing spontaneous photon emission into the fluctuating quantum-electrodynamic vacuum is an emblematic example of the dynamics of an open quantum system. Recent experiments have demonstrated that the gapped photon dispersion in periodic structures, which prevents photons in certain frequency ranges from propagating, can give rise to unusual spontaneous-decay behaviour, including the formation of dissipative bound states[1–3]. So far, these effects have been restricted to the optical domain. Here we demonstrate similar behaviour in a system of artificial emitters, realized using ultracold atoms in an optical lattice, which decay by emitting matter-wave, rather than optical, radiation into free space. By controlling vacuum coupling and the excitation energy, we directly observe exponential and partly reversible non-Markovian dynamics and detect a tunable bound state that contains evanescent matter waves. Our system provides a flexible platform for simulating open-system quantum electrodynamics and for studying dissipative many-body physics with ultracold atoms[4–6].

The Weisskopf–Wigner model of spontaneous emission[7,8], a central concept in quantum optics[9], describes how an excited atom can decay to its ground state owing to coupling to zero-point oscillations of the electromagnetic vacuum. It simultaneously represents one of the first open quantum systems discussed in the literature and an area of research that has recently seen a resurgence of intense theoretical efforts[10–13]. In its usual Markovian formulation, the model makes the assumption that the decay proceeds on a much slower timescale than the optical period, which leads to a memoryless, exponential decay of the excited-state amplitude and to an associated Lamb shift of the transition frequency. For free-space emission, the Markovian approximation is generally fulfilled to high accuracy.

On the other hand, modifications to the mode density of the vacuum can change the features of spontaneous decay. This was recognized first in the 1940s[14] and again decades later[15], during the development of cavity quantum electrodynamics[16–18], where the decay can be altered to the extreme point of coherent vacuum Rabi oscillations when the spectrum is restricted to a single mode. Between these two limits lies the regime of a vacuum with a bounded continuous spectrum, in which a strong modification of spontaneous decay behaviour occurs close to the boundary. An example is photonic crystals (also called photonic-bandgap materials)[19,20], where a periodic spatial modulation of the refractive index gives rise to a gapped dispersion relation. For emission close to a bandgap, the Markovian approximation can no longer be applied, and novel features appear, including oscillatory decay dynamics for energies above the band edge and the formation of atom–photon bound states below it[21]. Over the past two decades, experiments on spontaneous emission in photonic-bandgap materials, including the microwave domain, have observed some of these effects, specifically modified spontaneous emission rates[22,23] and Lamb shifts[24], as well as spectral signatures for non-exponential decay[1]. Very recently, experiments have probed the long-predicted atom–photon bound state[25,26], using both transmon qubits coupled to corrugated microwave guides[3] and atoms in photonic-crystal waveguides[2], with the prospect of engineering systems with optical long-range interactions[27].

Here, we realize an atom–optical analogue[4–6] of emission in a one-dimensional photonic-bandgap material, where the singularity in the mode density near the edge of the continuum leads to particularly strong deviations from Markovian behaviour. In our system of matter-wave emitters, the free tunability of the excitation energy and decay strength allows for a systematic exploration of the emergence of non-Markovian dynamics, including partial reversibility and the formation of a matter-wave bound state that can be directly detected. Importantly, the close spacing of emitters gives rise to collectively enhanced dynamics beyond the Weisskopf–Wigner model.

The experimental configuration is shown in Fig. 1a. Using a deep three-dimensional optical lattice with state selectivity along one axis, we prepare a sparse array of atoms confined to sites that are embedded in a system of isolated tubes acting as one-dimensional waveguides (see Methods for details). An atom's internal state ($|r\rangle$, red) is coherently coupled to a second, unconfined internal state ($|b\rangle$, blue) using an oscillatory magnetic field. Each site thus acts as a two-level matter-wave emitter, with harmonic-oscillator ground-state occupational levels $|g\rangle$ (empty) and $|e\rangle$ (full), supporting both the emission (for $|e\rangle \rightarrow |g\rangle$) and the absorption (for $|g\rangle \rightarrow |e\rangle$) of a $|b\rangle$ atom. The excitation energy of the emitter, which is given by the detuning, Δ, of the coherent coupling from the atomic resonance, is converted into kinetic energy for atomic motion along the axis of the waveguide.

One of the main features of each matter-wave emitter is its ability to undergo spontaneous decay, as described by the Weisskopf–Wigner model. Assuming no lattice potential, the driven atom performs simple Rabi oscillations between two internal states $|r\rangle$ and $|b\rangle$. These oscillations are described by the Hamiltonian $\hat{H} = (\hbar\Omega/2)e^{i\delta t}\hat{r}\hat{b}^{\dagger} + \text{h.c.}$, where Ω denotes the strength and δ the detuning of the coupling from the bare atomic resonance, $\hbar$ is the Planck constant and 'h.c.' denotes the Hermitian conjugate. The tight confinement of just one of the states (here, $|r\rangle$) strongly couples the atom's internal and motional degrees of freedom, producing a zero-point energy shift of $\bar{\varepsilon}_0 = \hbar\omega_0/2 \gg \hbar\Omega$, where ω_0 is the harmonic-oscillator frequency in the potential, as well as a kinetic-energy shift of $\varepsilon_k = \hbar^2k^2/(2m)$ for the motion of the free $|b\rangle$ state at $\hbar k$ momentum. As a consequence, the detuning and strength of the coupling are shifted to $\Delta_k = \delta + (\bar{\varepsilon}_0 - \varepsilon_k)/\hbar$ and $\Omega_k = \Omega\gamma_k$, respectively, with $\gamma_k = \langle k|\psi_e\rangle$ denoting the overlap of the external wavefunctions. Integration over all possible momenta k then yields[6] $\hat{H} = \sum_k \hbar g_k e^{i\Delta_k t}|g\rangle\langle e|\hat{b}_k^{\dagger} + \text{h.c.}$, with $g_k = \Omega_k/2$; that is, the standard Weisskopf–Wigner Hamiltonian describing spontaneous emission into a vacuum of modes (k, ε_k) (see Fig. 1b). In contrast to optical emission in free space, the dispersion relation ε_k is quadratic, as in a photonic crystal (see Fig. 1c). In such crystals, the emission energy relative to the edge of the continuum may be adjusted through the crystal's band structure; in our system, the excitation energy, $\hbar\Delta \equiv \hbar\Delta_{k=0}$, itself is tunable, including the case $\Delta < 0$. Importantly, the tunability also includes the vacuum coupling, g_k, which is set by Ω.

A common scenario considered in the Wigner–Weisskopf model is emission deep into the continuum, such that the decay dynamics is

[1]Department of Physics and Astronomy, Stony Brook University, Stony Brook, NY, USA. *e-mail: dominik.schneble@stonybrook.edu

© 2018 Springer Nature Limited. All rights reserved.

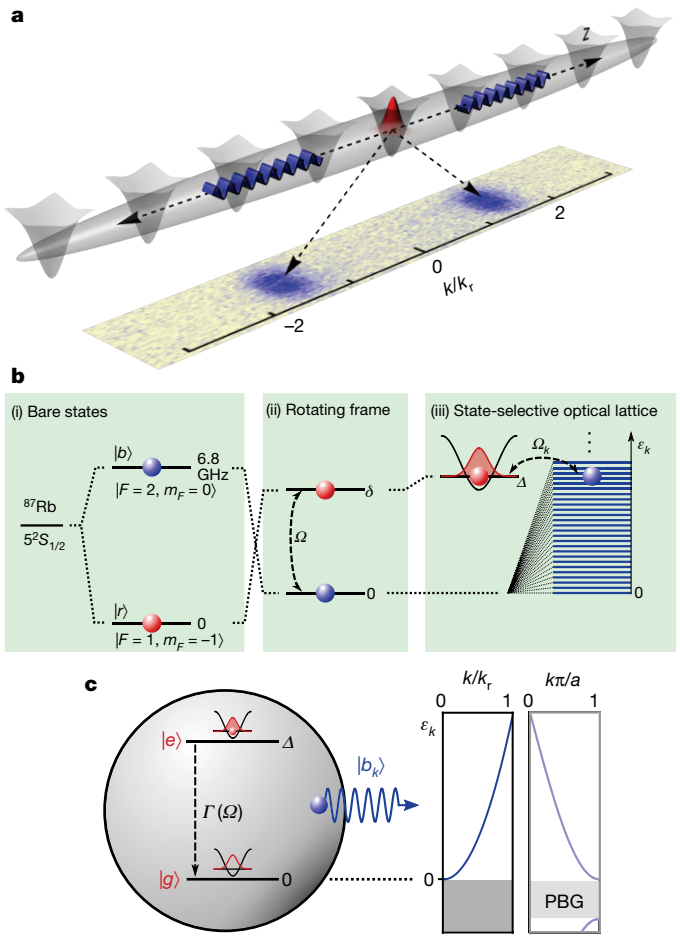


Fig. 1 | Realization of matter-wave emitters. a, Experimental configuration. An occupied site of an optical lattice embedded in a single-mode matter waveguide acts as an elementary emitter of a single atom; adjacent empty lattice sites act as absorbers. The bottom illustration shows the momentum distribution in the waveguide after release and free expansion, where $\hbar k_r = \hbar(2\pi/\lambda_z)$ (with a wavelength of $\lambda_z = 790.1$ nm) is the recoil momentum. **b**, Emission mechanism: (i) bare internal-state pair in ^{87}Rb (hyperfine ground states $|r\rangle = |F=1, m_F=-1\rangle$ and $|b\rangle = |F=2, m_F=0\rangle$ split by 6.8 GHz); (ii) state pair in a frame co-rotating with a near-resonant microwave field with variable detuning δ and coupling strength Ω; and (iii) state pair in the co-rotating frame after applying the state-selective lattice potential (detuning shifted to $\Delta = \delta + \omega_0/2$, where $\omega_0 = h \times 40$ kHz). The microwave couples (with $\Omega_k = \Omega\gamma_k$) the trapped $|r\rangle$ state to free $|b_k\rangle$ states with momentum k and kinetic energy ε_k. **c**, The filled (empty) potential well can be viewed as the excited (ground) state $|e\rangle$ ($|g\rangle$) of a matter-wave emitter. The emission of atoms in this scenario is similar to the emission of photons in photonic-bandgap (PBG) materials, with both featuring quadratic dispersions and energetically forbidden regions.

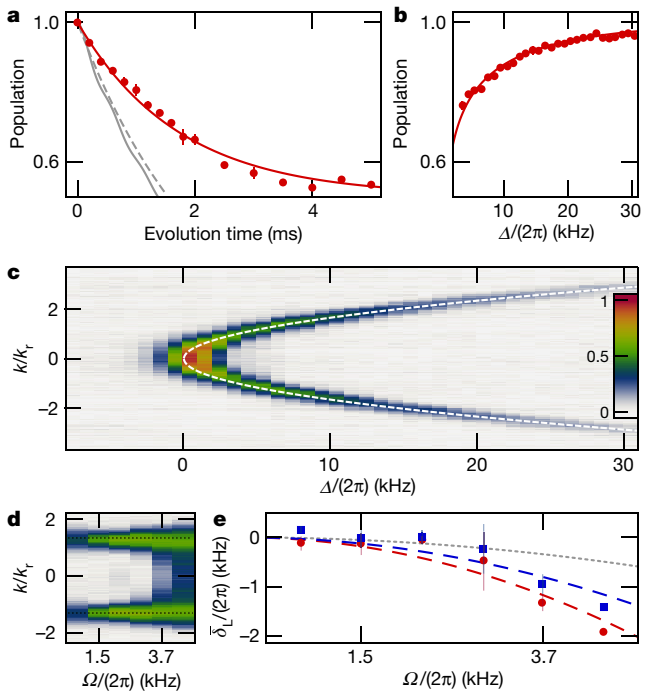


Fig. 2 | Markovian regime. a, Time evolution of the lattice population for $\Omega = 2\pi \times 0.74(5)$ kHz and $\Delta = 2\pi \times 1.9(3)$ kHz (red symbols). Each point is the average of at least three measurements and error bars show the standard error of the mean (s.e.m.). The red line is a phenomenological exponential decay curve with a fitted rate of $2\pi \times 94(3)$ Hz and an offset of 0.503(4). The light-grey lines represent the Markovian approximation (dashed, $\Gamma = 2\pi \times 72(12)$ Hz) and the exact analytical solution for an isolated emitter[6] (solid). **b**, Lattice population as a function of Δ for $t = 0.4$ ms and $\Omega = 2\pi \times 1.5(1)$ kHz. The solid line is the Markovian expectation with the overall decay width, Γ, scaled by 0.61(1). **c**, Detected momentum distribution of $|b\rangle$ atoms versus Δ for parameters as in **b**. The dashed line is the single-particle dispersion; data for small positive and for negative detunings are outside the Markovian regime. **d**, Raw time-of-flight data for extracting the energy shift at $\Delta = 2\pi \times 6.0(3)$ kHz. Colour scale as in **c**. **e**, Measured shifts $\bar{\delta}_L = \Delta_{\bar{k}} - \Delta$ in the regime $\Omega/\Delta < 1$ for $\tilde{\Omega}t = 1.24$ and averaged over $\Delta = 2\pi \times \{1, 2, 4, 6\}$ kHz. The data are extracted from the second moment (maximum) of the momentum distribution, shown by blue squares (red circles). Error bars show the s.e.m. The blue solid and red dashed lines are quadratic fits and the grey dotted line represents $\delta_L(\Omega)$.

much slower than the timescale set by the excited-state energy (or the elevation above the band edge in the case of a photonic crystal). This allows for a Markovian treatment and results in exponential decay of the excited state. Following Fermi's golden rule, the decay width, Γ, is the product of the mode density, ρ, and the square of a matrix element, H_{ge}, which for optical decay is the product of the electric dipole moment and the zero-point field of the resonant mode. For our system, an analogous analysis[6] (valid for $\Omega/\Delta \ll 1$) leads to $\Gamma = \Omega_{\bar{k}}^2/\sqrt{\omega_0\Delta}$, containing the one-dimensional mode density $\rho \propto 1/\sqrt{\Delta}$ and $H_{ge} \propto \hbar\Omega_{\bar{k}}$, where $\bar{k} = \sqrt{2m\Delta/\hbar}$ represents the resonant mode.

Because of the residual axial tube confinement, ω_z, all measurements are taken for $\Omega/\omega_z > 1$ and associated timescales shorter than $\tau_z = 2\pi/\omega_z \approx 10$ ms. The measured $|r\rangle$ population is shown in Fig. 2 for parameters in the (quasi-)Markovian regime $(\Omega/\Delta)^2 \ll 1$ as a function of time (Fig. 2a) and detuning (Fig. 2b); the data in Fig. 2a are for $\Omega/\Delta \approx 0.4$ at $\Delta \approx 2\pi \times 2$ kHz (with $\Gamma = 2\pi \times 72$ Hz). After a variable coupling time t, we observe an irreversible, exponential decay in agreement with the expectation; however, the measured population does not decay to zero but instead saturates at a finite value. We qualitatively explain this discrepancy by taking into account that an excited emitter is not isolated but part of a (mostly) ground-state array that enables reabsorption, in analogy to an optically thick medium. In the Weisskopf–Wigner formalism, the array is modelled by introducing site-dependent phases and projectors, resulting in[4] $\hat{H} = \sum_{k,j} \hbar g_k e^{i(\Delta_k t - kz_j)} |g_j\rangle\langle e_j| \hat{b}_k^\dagger + \text{h.c.}$ Following a master-equation-based treatment[5], we expect excitations to be transferred between neighbours on a timescale of $t \approx 1/\Gamma = 2$ ms, leading to a slowdown of the decay, in qualitative agreement with Fig. 2a (see also Methods section 'Numerical simulation of array effects'). Moreover, the emitted atoms cannot escape from our system for long times, which induces the formation of a steady state as t approaches τ_z. Here, additional dephasing effects that are not considered in this model may arise from collisions between emitted atoms (scattering between modes in each tube). At early times, $t < 1/\Gamma$, with still weak reabsorption, the decay at a fixed time τ (see Fig. 2b) displays approximately the expected detuning dependence of the Weisskopf–Wigner model, $\exp[-\Gamma(\Delta)\tau] \approx 1 - \Gamma(\Delta)\tau$ (albeit with a downward

© 2018 Springer Nature Limited. All rights reserved.

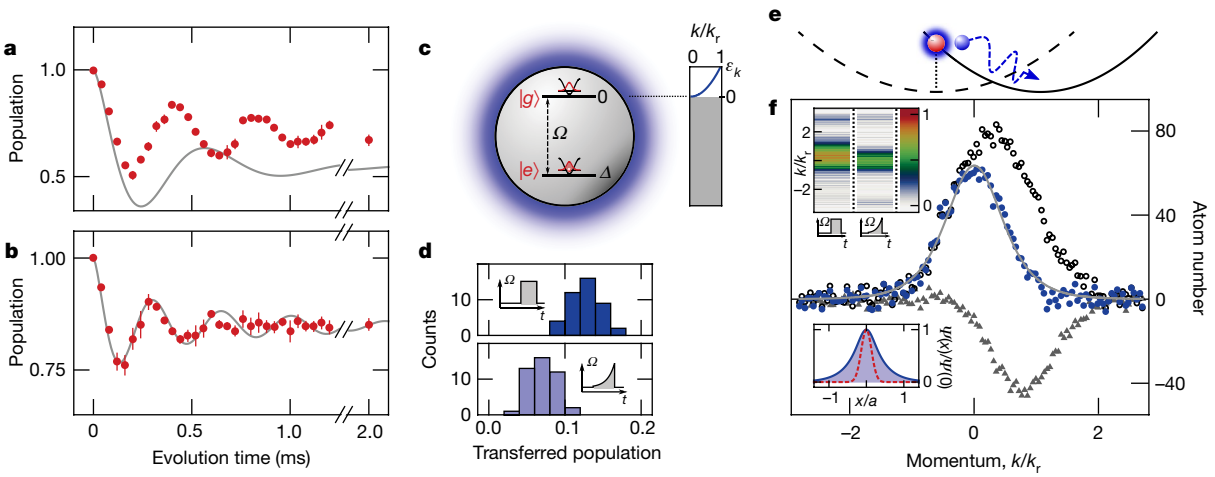


Fig. 3 | Non-Markovian dynamics and bound-state formation. **a**, Time evolution of the $|r\rangle$ population for $\Delta = -2\pi \times 0.1(3)$ kHz and $\Omega = 2\pi \times 3.0(3)$ kHz (with points and error bars as in Fig. 2a). The grey line is our analytical model. **b**, Same as **a**, but for $\Delta = -2\pi \times 1.7(3)$ kHz. The fitting parameters of the analytical model are $\Delta = -2\pi \times 2.08(3)$ kHz and $\Omega = 2\pi \times 2.79(4)$ kHz. **c**, Illustration of a stationary bound state for negative excitation energies. **d**, Asymptotic fraction of $|b\rangle$ atoms, extracted from the time evolution after $t = 2.6$ ms, for Ω as in **a** and $\Delta = -2\pi \times 2.2(3)$ kHz. The top (bottom) histogram refers to the case of a sudden (adiabatic) turn-on of the coupling, measured from 50 experimental runs each. **e**, Illustration of separation of evanescent and propagating waves along the weakly confining tube axis. The atomic sample resides off-centre from the minimum of the axial tube potential (solid line), shifted from that of the initial trap (dashed line), which defines the sample position. While the sample and bound state are frozen in place (vertical dotted line), the nonadiabatically released fraction (blue arrow) is accelerated towards the potential minimum. **f**, Momentum distributions of $|b\rangle$ atoms for the two scenarios considered in **d**. Open (filled) circles represent the sudden (adiabatic) turn-on of the coupling, averaged over all 50 runs, and triangles show the difference of the two datasets. The solid line represents the square of the Fourier transform of the analytical evanescent wavefunction[6], fitted to the adiabatic data ($\Delta_{\mathrm{fit}} = 2\pi \times 2.1(1)$ kHz). The centre of mass of the non-adiabatic dataset is shifted by $p_s = 0.32(1)\hbar k_r$ relative to the adiabatic dataset, and that of their difference (with a half-width at half-maximum of $0.57(1)\hbar k_r$) is shifted by $p_s = 0.80(1)\hbar k_r$. The expectation for free atoms is $p_s = 0.83(7)\hbar k_r$. Lower inset, corresponding real-space evanescent wavefunction (blue, solid) and Wannier function of a lattice-trapped $|r\rangle$ atom (dashed, red). Upper inset, raw data for the momentum distributions of **e**, before subtraction of spurious higher-band contributions (see Methods).

rescaling of the actual value of Γ at the finite pulse time $\tau = 0.4$ ms used, with reabsorption effects already apparent).

We next characterize the momentum distribution of the emitted atoms. For this purpose we apply a 0.4-ms-long coupling pulse and then observe the location of the $|b\rangle$ atoms after 15 ms of free fall, using state-selective absorptive imaging (see Methods). On the basis of the Markovian approximation, isotropic emission with wavepackets centred near the resonant momentum $\bar{k}(\Delta)$ is expected. In Fig. 2c we show the observed momentum distribution as a function of Δ; the emission clearly exhibits parabolic dispersion. Moreover, the spectral width, σ_k, of the separated wavepackets decreases with detuning, in qualitative agreement with the expectation ($\sigma_k \propto 1/\Delta$ for large detunings[6]; a quantitative comparison is compromised by the finite time of flight). The 'intensity' of the emitted matter-wave pulse strongly depends on the detuning, as shown in Fig. 2b.

The standard Markovian treatment of the Weisskopf–Wigner model yields a Lamb shift of the ground and excited states as a unitary coupling to the vacuum. An analogous analysis for our system[4,6] yields a shift $\delta_L = \Omega^2/\omega_0$ of the excited-state energy, to $(\Delta - \delta_L)$. We measure the momentum distribution for variable Ω at several values of Δ and then calculate the mean kinetic energy of the wavepackets both from the second moment of the momentum distributions and from the location of their fitted maxima (see Methods for details). To facilitate comparison with the model, the data are taken for a constant effective pulse area $\tilde{\Omega}t$, where $\tilde{\Omega} = (\Omega/\omega_0)^{1/3}\Omega$, and $t \leq 1$ ms to mitigate propagation effects. The results for the (quasi-)Markovian regime $(\Omega/\Delta)^2 \ll 1$ are shown in Fig. 2d, e as a function of Ω. The extracted shift has the sign and approximate quadratic dependence of δ_L but is a factor of roughly three larger. We caution that while this alone could point towards the existence of collective enhancement, there is no indication for superradiance[4,28] from the decay data (which is consistent with the fact that there is no overall population inversion in the array).

Our system readily allows the study of spontaneous emission outside the Markovian regime, as shown in Fig. 2c. In particular, the diverging one-dimensional mode density near $\varepsilon_k = 0$ greatly enhances the effects of the edge of the continuum. For emission at low excitation energy, $\Delta/\Omega \ll 1$, we expect dynamics reminiscent of a two-level system, with damping provided by low-energy modes. Results of measurements at $\Delta = 0$ are shown in Fig. 3a. We observe oscillations similar to the predictions of our isolated-emitter model[6] (which now features a finite offset, in contrast to positive detunings) but with higher frequency and less damping, suggesting that the dynamics is coherently enhanced by low-energy modes whose wavelengths can extend over several emitters (see also Methods section 'Numerical simulation of array effects').

For emission below the continuum edge (that is, for $\Delta < 0$), we expect the formation of a stationary bound state[4,6], as illustrated in Fig. 3c. For our one-dimensional system, this state consists of a partly excited emitter dressed by an evanescent, approximately exponentially decaying matter wave, with a binding energy of $\hbar\omega_B \approx \hbar\Delta$ and a localization length[6] of $\xi = 1/\sqrt{2m|\omega_B|/\hbar}$. To isolate the bound state, the coupling needs to be turned on slowly to prevent the additional population of freely propagating modes[6] representing a non-adiabatic, transient shedding of matter waves. However, first we proceed as before by switching on the coupling, at $\Delta = -2\pi \times 1.7$ kHz. The lattice population, shown in Fig. 3b, shows a transient oscillation (with much lower amplitude than at the edge) settling to an asymptotic value below unity (with an observable $|b\rangle$ population; see Fig. 2c). Remarkably, our isolated-emitter model[6] now closely fits the data within the experimental uncertainties. Indeed, for the chosen parameters, ξ is less than half a lattice period, which should lead to a relative suppression of long-range couplings.

To access the properties of the bound state, we first determine the fraction of $|b\rangle$ atoms by comparing the asymptotic lattice population for a sudden and for an exponential turn-on of the coupling. The results in Fig. 3d show that 7.1(2)% of the population are in the evanescent wave, with a total $|b\rangle$ population of 12.7(2)% (here and elsewhere the errors quoted denote one standard deviation). The observed fraction of $|b\rangle$ atoms in the bound state (55%) is close to the expectation[6] of 47% for the chosen parameters, with the excess possibly stemming from residual non-adiabaticity of the ramping (we noticed an inconsistency in ref. [6] for the total $|b\rangle$ fraction, but this does not affect the relative $|b\rangle$ fraction in the bound state). Importantly, the recorded momentum distribution of

© 2018 Springer Nature Limited. All rights reserved.

the bound evanescent fraction of $|b\rangle$ atoms (see Fig. 3f) can be directly compared to the prediction of our analytical model[6] owing to the absence of propagation effects. Fitting the model to the data for the adiabatic ramp-up with Δ as the only fitting parameter reproduces its experimental value to within the experimental uncertainty, and the agreement between the model and the measured momentum distribution is indeed excellent. The localization length obtained from the exact model is $\xi = 142(3)$ nm (we note that for the parameters chosen[6], $\omega_B \approx 1.3\Delta$).

Finally, additional direct evidence for the non-adiabatically released part can be obtained by comparing the adiabatic momentum distribution to that recorded after the rectangular pulse. Because our sample is prepared off-centre from the potential minimum along the tube direction (see Fig. 3e, Methods), the distribution of the shed $|b\rangle$ atoms can separate in momentum space from the bound fraction. The pulse duration used in this experiment corresponds to about a quarter of the oscillation period (2.6 ms) along the tube axis, thus maximizing this differential effect. As seen in Fig. 3f, the (symmetric) difference of the two momentum distributions is centred at the finite momentum expected for atoms released with zero velocity at the beginning of the pulse. This finding is consistent with the time evolution in Fig. 3b, which suggests that the release of the unbound fraction occurs within a short time (about 0.5 ms) after the coupling is turned on. We note that as a result of propagation in the tubes for a quarter of the oscillation period, the width of the momentum distribution of the released atoms reflects that of their distribution in real space ($v_z = z\omega_z$). The extracted half-width at half-maximum of 5.47(9) μm is comparable to the Thomas–Fermi radius of the initial condensate.

Much of the present work has focused on basic properties arising from the tunability of our Wigner–Weisskopf system, including the formation of bound states below the edge of the mode continuum. On the single-emitter level, this provides a direct analogy to atomic decay near the bandgap of a photonic crystal. We note that in yet another context, the observed non-Markovian oscillatory dynamics also reproduces predictions for electron photodetachment from negative ions[29,30]. The optical lattice geometry opens up various additional avenues of inquiry. For emission sufficiently above the continuum edge, these may include novel types of superradiance that depend on the degree of coherence of the lattice population (superfluid or Mott insulating)[4,5] and have no analogue in optical systems. Moreover, controlling the longitudinal waveguide level spacing should enable studies of the transition between Dicke- and Tavis–Cummings-type models in quantum optics[4,9] (restricted to co-rotating terms), including their modification in the non-Markovian regime. Unlike photons, the emitted atoms can directly interact with each other, which should give rise to additional, nonlinear effects that modify the population dynamics. For negative energies, the bound state lends itself to the realization of lattice models[4] with modified tunnelling and interactions. Superficially, the structure of the bound state resembles that of a lattice polaron[31,32] (for which a phononic Lamb shift has recently been measured[33]), with massive vacuum excitations replacing massless Bogoliubov sound excitations. Rather than reducing transport, the bound state here leads to an enhancement of mobility. The presence of tunnelling with a tunable range is of interest, for example, for studies of integrability and thermalization in one-dimensional geometries.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0348-z.

Received: 20 December 2017; Accepted: 25 May 2018;
Published online 25 July 2018.

1. Hoeppe, U. et al. Direct observation of non-Markovian radiation dynamics in 3D bulk photonic crystals. *Phys. Rev. Lett.* **108**, 043603 (2012).
2. Hood, J. D. et al. Atom–atom interactions around the band edge of a photonic crystal waveguide. *Proc. Natl Acad. Sci. USA* **113**, 10507–10512 (2016).
3. Liu, Y. & Houck, A. A. Quantum electrodynamics near a photonic bandgap. *Nat. Phys.* **13**, 48–52 (2017).
4. de Vega, I., Porras, D. & Cirac, J. I. Matter-wave emission in optical lattices: single particle and collective effects. *Phys. Rev. Lett.* **101**, 260404 (2008).
5. Navarrete-Benlloch, C., de Vega, I., Porras, D. & Cirac, J. I. Simulating quantum-optical phenomena with cold atoms in optical lattices. *New J. Phys.* **13**, 023024 (2011).
6. Stewart, M., Krinner, L., Pazmiño, A. & Schneble, D. Analysis of non-Markovian coupling of a lattice-trapped atom to free space. *Phys. Rev. A* **95**, 013626 (2017).
7. Weisskopf, V. & Wigner, E. Berechnung der natürlichen Linienbreite auf Grund der Diracschen Lichttheorie. *Z. Phys.* **63**, 54–73 (1930).
8. Milonni, P. W. *The Quantum Vacuum: An Introduction to Quantum Electrodynamics* (Academic Press, San Diego, 1994).
9. Meystre, P. & Sargent, M. III *Elements of Quantum Optics* (Springer, Berlin, 2007).
10. Rivas, A., Huelga, S. F. & Plenio, M. B. Quantum non-Markovianity: characterization, quantification and detection. *Rep. Prog. Phys.* **77**, 094001 (2014).
11. Rotter, I. & Bird, J. P. A review of progress in the physics of open quantum systems: theory and experiment. *Rep. Prog. Phys.* **78**, 114001 (2015).
12. Breuer, H.-P., Laine, E.-M., Piilo, J. & Vacchini, B. Non-Markovian dynamics in open quantum systems. *Rev. Mod. Phys.* **88**, 021002 (2016).
13. de Vega, I. & Alonso, D. Dynamics of non-Markovian open quantum systems. *Rev. Mod. Phys.* **89**, 015001 (2017).
14. Purcell, E. M. Spontaneous emission probabilities at radio frequencies. *Phys. Rev.* **69**, 681 (1946).
15. Kleppner, D. Inhibited spontaneous emission. *Phys. Rev. Lett.* **47**, 233–236 (1981).
16. Miller, R. et al. Trapped atoms in cavity QED: coupling quantized light and matter. *J. Phys. B* **38**, S551–S565 (2005).
17. Walther, H., Varcoe, B. T. H., Englert, B.-G. & Becker, T. Cavity quantum electrodynamics. *Rep. Prog. Phys.* **69**, 1325–1382 (2006).
18. Haroche, S. & Raimond, J.-M. *Exploring the Quantum: Atoms, Cavities, and Photons* (Oxford Univ. Press, Oxford, 2006).
19. John, S. Strong localization of photons in certain disordered dielectric superlattices. *Phys. Rev. Lett.* **58**, 2486–2489 (1987).
20. Yablonovitch, E. Inhibited spontaneous emission in solid-state physics and electronics. *Phys. Rev. Lett.* **58**, 2059–2062 (1987).
21. Lambropoulos, P., Nikolopoulos, G. M., Nielsen, T. R. & Bay, S. Fundamental quantum optics in structured reservoirs. *Rep. Prog. Phys.* **63**, 455–503 (2000).
22. Tocci, M. D., Scalora, M., Bloemer, M. J., Dowling, J. P. & Bowden, C. M. Measurement of spontaneous-emission enhancement near the one-dimensional photonic band edge of semiconductor heterostructures. *Phys. Rev. A* **53**, 2799–2803 (1996).
23. Lodahl, P. et al. Controlling the dynamics of spontaneous emission from quantum dots by photonic crystals. *Nature* **430**, 654–657 (2004).
24. Liu, Q. et al. Observation of Lamb shift and modified spontaneous emission dynamics in the YBO_3: Eu^{3+} inverse opal. *Opt. Lett.* **35**, 2898–2900 (2010).
25. Bykov, V. P. Spontaneous emission from a medium with a band spectrum. *Sov. J. Quantum Electron.* **4**, 861–871 (1975).
26. John, S. & Wang, J. Quantum electrodynamics near a photonic band gap: photon bound states and dressed atoms. *Phys. Rev. Lett.* **64**, 2418–2421 (1990).
27. Douglas, J. S., Habibian, H., Hung, C.-L., Gorshkov, A. V., Kimble, H. J. & Chang, D. E. Quantum many-body models with cold atoms coupled to photonic crystals. *Nat. Photon.* **9**, 326–331 (2015).
28. Dicke, R. H. Coherence in spontaneous radiation processes. *Phys. Rev.* **93**, 99–110 (1954).
29. Rzążewski, K., Lewenstein, M. & Eberly, J. H. Threshold effects in strong-field photodetachment. *J. Phys. B* **15**, L661–L667 (1982).
30. Lewenstein, M. & Rzążewski, K. Quantum anti-Zeno effect. *Phys. Rev. A* **61**, 022105 (2000).
31. Bruderer, M., Klein, A. Clark, S. R. & Jaksch, D. Polaron physics in optical lattices. *Phys. Rev. A* **76**, 011605, 2007.
32. Gadway, B., Pertot, D., Reimann, R. & Schneble, D. Superfluidity of interacting bosonic mixtures in optical lattices. *Phys. Rev. Lett.* **105**, 045303 (2010).
33. Rentrop, T. et al. Observation of the phononic Lamb shift with a synthetic vacuum. *Phys. Rev. X* **6**, 041041 (2016).

Acknowledgements We thank M. G. Cohen for discussions and a critical reading of the manuscript. This work was supported by NSF PHY-1607633. M.S. was supported by a GAANN fellowship by the US Department of Education. A.P. acknowledges partial support from ESPOL-SENESCYT.

Author contributions D.S., L.K. and M.S. conceived the experiment. L.K. took the measurements with assistance from A.P. and J.K. L.K. analysed the data with contributions from M.S. Numerical simulations were performed by L.K. D.S. supervised the project. The results were discussed and interpreted by all authors. The manuscript was written by L.K. and D.S. with contributions from A.P., J.K. and M.S.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0348-z.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to D.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Experimental procedures. In this section we outline experimental procedures implemented in our measurements.

Sample preparation. In the experiment we used ^{87}Rb atoms in the hyperfine ground states $|r\rangle = |F=1, m_F=-1\rangle$ and $|b\rangle = |2, 0\rangle$ (the fact that $|r\rangle$ lies below $|b\rangle$ is inconsequential in the rotating frame). The atoms are confined to a two-dimensional array of roughly 10^3 isolated lattice tubes spaced at 532 nm, each with a radial confinement of $\omega_\perp = 2\pi \times 26$ kHz and a residual axial confinement of $\omega_z = 2\pi \times 97$ Hz that quantizes the mode spectrum for released $|b\rangle$ atoms in the z direction; this confinement is inconsequential for times much shorter than $\tau_z = 2\pi/\omega_z \approx 10$ ms. A state-selective lattice with period $\lambda_z/2 = 395$ nm and harmonic-oscillator frequency $\omega_0 = 2\pi \times 40(1)$ kHz strongly confines the $|r\rangle$ atoms along the tube axis.

Starting with an optically trapped Bose–Einstein condensate[34], we first create an atomic Mott insulator of $|r\rangle$ atoms by simultaneously ramping up all three optical lattices over a time of 90 ms to depths of $40E_{r,1,064nm}$,$40E_{r,1,064nm}$ and $30E_{r,790.1nm}$, where $E_{r,\lambda} = (h/\lambda)^2/(2m)$ is the respective recoil energy, in the x, y and z directions, respectively. This procedure results in a deeply confined Mott-insulating sample, while also shifting the minimum of the tube potentials by about 10 μm with respect to the sample. The shift (which is comparable to the sample size and corresponds to the change of the gravitational sag during the ramp[34] after the atoms are effectively pinned by the z lattice) leaves the resonance condition for individual atoms unaffected because $|r\rangle$ and $|b\rangle$ experience the same tube potential.

After loading the atoms into the lattice, we transfer a fraction of approximately 0.82 to an intermediate state $|2, 1\rangle$ using a two-photon radiofrequency pulse (microwave radiation of 6.8 GHz and radiofrequency radiation of about 3 MHz) of 0.95 ms duration. These atoms are eventually removed using resonant light on the D_2 cycling transition ($F=2 \rightarrow F'=3$). Using the $|2, 1\rangle$ state has the advantage that there is no first-order shift in the energy difference between $|1, -1\rangle$ and $|2, 1\rangle$ due to either magnetic field or the state-selective optical potential. Also, collisional shifts[35] of doubly (triply) occupied sites are roughly 50 Hz (100 Hz) for our lattice parameters, much less than the spectral width of the pulse. This yields a sample of $2.8(2) \times 10^4$ $|r\rangle$ atoms with an average site occupation of $\langle n_i \rangle \lesssim 0.5$ in the tubes. Having thus created an initial state of matter-wave emitters, we then switch on a 6.8-GHz microwave field of variable coupling strength Ω and detuning Δ.

Atom detection. We switch off the microwave coupling Ω suddenly, and we measure the population remaining in the lattice and access the momentum distribution of the released atoms with state-selective absorptive imaging, using a combination of band mapping and Stern–Gerlach separation during the time of flight (TOF).

The state detection begins with a 500-μs bandmap step during which all lattice potentials are ramped down to zero (the bandmap step prevents a rapid transverse expansion out of the tube potentials and is therefore favourable for momentum-space detection when small atom numbers are used). Subsequently all remaining trapping potentials are turned off suddenly, and after an expansion time of about 1 ms a pulse sequence is applied for magnetic-field characterization (see below), immediately followed by a 5-ms-long magnetic field gradient pulse to separate states of different magnetic moments. Finally, 14.5 ms after the release from the optical potential, a 200-μs-long imaging pulse of resonant light on the $F=2 \rightarrow F'=3$ D_2 cycling transition is used to detect the $F=2$ atoms; this yields a total effective TOF of 14.6 ms for atoms in this state. The $F=1$ atoms are repumped for 100 μs using resonant D_2 light ($F=1 \rightarrow F'=2$) after an additional 2.7 ms, immediately followed by another 200-μs-long imaging pulse (D_2, $F=2 \rightarrow F'=3$) to detect the repumped $F=1$ atoms after a total TOF of 17.6 ms. Because all magnetic moments are unique within $F=1$ and $F=2$ separately, we fully resolve the population in each individual hyperfine state $|F, m_F\rangle$. We note that technically the cloud centres of $F=1$ and $F=2$ along the imaging direction are still overlapped during the second imaging pulse, however the $F=2$ atoms are pushed out of the field of view by the time the $F=1$ detection occurs.

Image analysis. Our raw data are images acquired using a Princeton Instruments PIXIS:1024B charge-coupled device (CCD) camera with WinView32 software in .spe format. We use Mathematica to read the raw image data and create standard absorption images. Residual fringes are removed by subtracting typical fringes found using principal-component analysis of a large set (>100) of empty absorption images.

Momentum-space calibration. Our standard momentum calibration relies on Kapitza–Dirac diffraction[36] from the z lattice. For a more precise determination of emission momenta, we take into account residual propagation in the tubes that slows the atomic motion. After ramping up the z lattice, the tubes are created by partial retro-reflection of the Gaussian beams of our optical trap ($1/e^2$ radius of $w = 135$ μm)[34], which leads to an increase of the optical confinement $\omega_z/(2\pi)$ from 72(1) Hz to 97(1) Hz (with the gravity direction along z). The tubes are again ramped down within 500 μs after the microwave pulse (together with the z lattice, for band-mapping purposes), followed by a switch-off of the optical trap. We numerically simulate the motion of atoms in the tubes by assuming that the release (with momentum $\pm\hbar k$) occurs midway through the pulse, at the centre of the 72-Hz trap, and by calculating the trajectories in the time-dependent optical potential until detection, after a TOF of 14.6 ms. We see that the calibration differs from the Kapitza–Dirac diffraction results by −0.5%, −1% and −6% for pulse durations of 0.2 ms, 0.4 ms and 1 ms, respectively, with negligible differences for shorter pulses. These corrections are included in Fig. 2.

Characterization of magnetic fields and optical potentials. The quoted uncertainty of 300 Hz in the detuning Δ has contributions from both differential Zeeman and alternating-current Stark shifts, which are characterized as follows.

Magnetic fields. All experiments are carried out at fields between 4.9998(1) G and 5.0002(1) G, where the error corresponds to an uncertainty in the bare-level splitting of 70 Hz between $|r\rangle$ and $|b\rangle$. The magnetic field for each iteration of the experiment is monitored using a series of Rabi pulses during the TOF[37], in which atoms in $|1, -1\rangle$ are redistributed to the $|2, -2\rangle$ and $|2, -1\rangle$ states. This method allows field reconstruction to within 100 μG (for details see ref. [37]). The inhomogeneity of the magnetic field across the sample is characterized using a Ramsey pulse sequence. This inhomogeneity does not exceed 70 μG, which corresponds to a root-mean-square (r.m.s.) variation of Δ of 15 Hz across the sample for the $(|r\rangle, |b\rangle)$ state pair.

Differential optical potential. The state-selective optical potential is created using a σ^--polarized laser beam (waist, 230 μm) at $\lambda_z = 790.10(2)$ nm. The polarization and wavelength are set (using a $\lambda/4$ waveplate and laser controls) such that the lattice potential seen by the $|r\rangle$ atoms is maximized, while the $|b\rangle$ atoms experience zero potential. The change in lattice potential with wavelength is $0.12E_r/(0.01\text{ nm})$, whereas the maximum theoretically possible polarization change (σ^- to π) changes the potential by $0.01E_r$ (we note that the polarization is stable in the experiment). We characterize the state selectivity using a sequence of ten Kapitza–Dirac pulses (10 μs) spaced at the Talbot resonance time[38], $\tau = (4E_{r,790.1nm}/h)^{-1} = 68$ μs, to ensure a suppression of the optical potential for $|b\rangle$ atoms by a factor of more than 100 with respect to that for $|r\rangle$ atoms, consistent also with the remaining wavelength uncertainty of 0.02 nm.

A crucial part of the experiment is the reliable determination of the resonance condition, or the value of Δ. Because the creation of our matter-wave emitter (Fig. 1c) starts with a detuned Rabi oscillation of two hyperfine states, we first (post-)stabilize the magnetic field as discussed above and in ref. [37]. The resonance condition ('excited-state energy') of the matter-wave emitters is determined by the detuning of the hyperfine spin and the zero point shifts to both states induced by the state-selective potential. Because we can currently control the wavelength only to an accuracy of 0.02 nm r.m.s., the overall uncertainty is limited to a maximum of ±340 Hz r.m.s., with the most extreme fluctuations limited to 1 kHz. To address and monitor this issue, we bracket each measurement (such as measuring the time evolution at a given detuning) by a resonance curve; an average of such resonance curves is shown in Extended Data Fig. 1 (see also figure 5 in ref. [37]). The resonance curve is taken without any transverse lattices on, with a low atom number in the optical trap (around 30,000 atoms) and only at partial transfer (at most 30%) to minimize systematic mean-field and density shifts. The residual systematic shift of the resonance condition due to mean-field and density effects is estimated to be less than 100 Hz by using direct simulation of resonance curves with the one-dimensional time-dependent Gross–Pitaevskii equation. The scatter of the centre of the resonance curves during a typical measurement has an r.m.s. value of 0.3 kHz, which is the quoted uncertainty of the detuning and also matches the expected wavelength reproducibility of 0.02 nm.

To characterize inhomogeneities of the state-selective optical potential, first we precisely calibrate the Rabi coupling strength (for a range of coupling strengths) in the absence of state-selective potentials. We then compare the maximum observed population transfer in a Rabi spectrum to the expected maximum population transfer into the lattice. Based on the comparison, we estimate an upper bound for the inhomogeneity of 300 Hz r.m.s. We note that the inhomogeneities of the trapping potential at 1,064 nm do not exceed 20 Hz r.m.s. across the sample. For comparison, the absolute magnitude of the spectroscopic shift of $|1, -1\rangle \leftrightarrow |2, 0\rangle$ in the centre of the optical trap does not exceed 400 Hz.

Background subtraction. The sequence used to thin out the atomic sample leaves roughly 10^3 atoms in the $|b\rangle$ state before the microwave pulse is applied. This results in a diffuse background in the momentum distributions, as illustrated in Extended Data Fig. 2. We remove this background by subtracting reference data taken for zero pulse time. The result is shown in Fig. 2c.

In Fig. 3f we also show the subtraction of background due to the very small fraction of higher-band population. The first excited band is expected at approximately $2\pi \times 36(1)$ kHz or $\pm 3\hbar k_r$. For the subtraction, we first fit three Gaussian peaks to the adiabatic ramp and one soft-edge box and two Gaussians to the sudden turn-on data, requiring identical atom numbers in the soft-edge box and the small Gaussian. The fitted functions are subtracted from the data, which then yield the main plot in Fig. 3f. The number of atoms emitted from higher bands is of the order of 510(30) for the sudden turn-on and 260(30) for the adiabatic ramp, which is less than 1 atom in a higher band per tube. The total number of atoms emitted

© 2018 Springer Nature Limited. All rights reserved.

from higher bands changes with the pulse area because spurious emission from the higher bands is in the Markovian regime.

Energy-shift data. The main motivation for the precise momentum calibration described above lies in the small size of the energy shift. Another challenge is the blurring of the distribution due to propagation effects for small coupling strengths (that is, long pulses). We use two measures for the determination of the energy of the emitted wavepackets: the squared separation of the wavepacket centres (extracted from the fit) and the second moment of the (centred) distribution. The accuracy of the peak-separation measure is limited by the fact that it ignores the physical broadening of the momentum distribution at larger coupling strengths and shorter times, whereas the second moment is sensitive to blurring of the wavepackets during detection. The data obtained using both methods are shown in Extended Data Fig. 3. In the non-Markovian regime, $\Omega/\Delta > 1$, the peaks become indistinguishable (see the apex of the parabola in Fig. 2c) and a meaningful measure of the shift cannot be extracted with either method.

Numerical simulation of array effects. As stated in the main text, the Hamiltonian of the multi-site array[4] predicts the (resonant) transfer of excitations between neighbouring emitters[5], in some similarity to the mechanisms for radiation trapping in an optically thick sample. Instead of attempting to solve this Hamiltonian, here we consider a simplistic model of an array of three emitters that is coupled to a quantized mode structure reflecting the weak longitudinal harmonic confinement. We chose our simplistic three-site model to have a centrally occupied site as the simplest spatially symmetric generalization of the isolated-emitter situation, which also corresponds to an 'average segment' in our system, with approximately one empty neighbour on each side of a populated site.

We start from the Rabi Hamiltonian (in the rotating-wave approximation) and expand it to couple one or several sites ($|r\rangle$) to many different, weakly confined levels ($|b\rangle$). This Hamiltonian is (for simplicity only shown for two sites, but readily expanded to n sites)

$$\hat{H} = \frac{\hbar}{2}\begin{bmatrix} 2\delta_1 & 0 & \Omega\gamma_{1,1} & \Omega\gamma_{1,2} & \cdots \\ 0 & 2\delta_2 & \Omega\gamma_{2,1} & \Omega\gamma_{2,2} & \\ \Omega\gamma_{1,1} & \Omega\gamma_{2,1} & -2\Delta+\omega_z & 0 & \\ \Omega\gamma_{1,2} & \Omega\gamma_{2,2} & 0 & -2\Delta+3\omega_z & \\ \vdots & & & & \ddots \end{bmatrix}$$

where $\delta_i = m\omega_z^2 r_i^2/(2\hbar)$ is a site-dependent detuning (that is, a site-dependent offset due to the weak harmonic confinement ω_z experienced by both lattice-trapped and free atoms) and $\gamma_{i,j}$ represents overlaps between final- and initial-state wavefunctions (calculated numerically). We use modes up to a fixed frequency ($\omega_{\max} = 2\pi \times 5$ kHz) and restrict ourselves to $\Delta + 2\Omega < \omega_{\max}$.

The results of the simulation for the Markovian parameters discussed in the main text are shown in Extended Data Fig. 4a, b. The simulated decay for an isolated emitter reproduces the prediction of our analytical model[6]—small discrepancies arise from the fact that the latter neglects terms of order $(\Delta/\omega_0)^2$ and $(\Omega/\omega_0)^2$ and higher, whereas our numerical simulation retains all orders. For the three-site array, the presence of neighbouring wells leads to a slowdown of the decay. The origin of this behaviour is reabsorption of emitted population by initially empty neighbours, as seen in Extended Data Fig. 4b. For the array considered, the process also leads to the formation of a temporary plateau in the overall site population, which may be related to the offset observed in the experiment. However, we caution that the long-time decay occurs in a regime not accessible experimentally (we assumed a denser mode structure to extend the continuum approximation; see below), and also the dynamics may be different in an optical lattice extending over the entire mode volume.

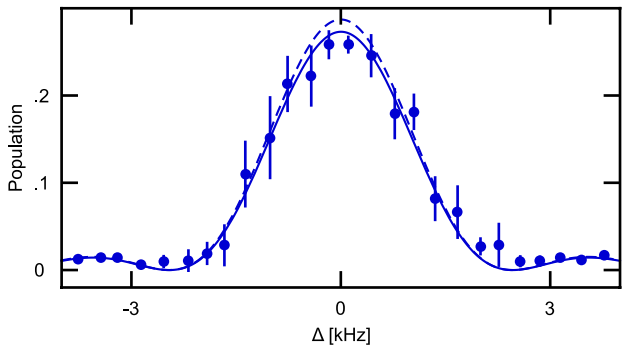
In Extended Data Fig. 4c we plot the dynamics in the extreme non-Markovian regime at the edge, where the coupling strength is much larger than the excited-state energy. In this case, we see that the single emitter displays oscillatory dynamics that quickly damps out and settles to a non-zero value, again in agreement with the analytical theory. On the other hand, the three-emitter array shows oscillatory dynamics of much greater amplitude and longer duration. We interpret this as a coherent enhancement of the dynamics through tunnelling to nearest neighbours. The range of this tunnelling diverges at the band edge, causing the marked difference between single emitter and three emitters.

From our simulations we gain additional insight into the effects of the quantized mode structure. As discussed in the main text, the mode structure should act like a true continuum for short enough times, where uncertainty should 'wash out' the levels. The simulations provide a quantitative test for this. The simplest comparison is for an isolated emitter in the Markovian limit. We see that for early times, the Markovian prediction[6] quantitatively agrees with the numerical solution, with a marked deviation ('revival') observable only at $t^* > 0.25[\omega_z/(2\pi)]^{-1}$ (we restrict data taking to $t < t^* = 2.6$ ms in the experiments, in all but one case.) This is independent of the set harmonic trapping frequency and the location of the site in the array (centre or off-centre). We note that similar results are also obtained if the continuum is discretized by assuming a periodic-box-type potential, where the revival time depends on the length of the box. We furthermore note that our simulation does not reproduce δ_L for long times.

Data availability. The data that support the findings of this study are available from the corresponding author upon reasonable request.

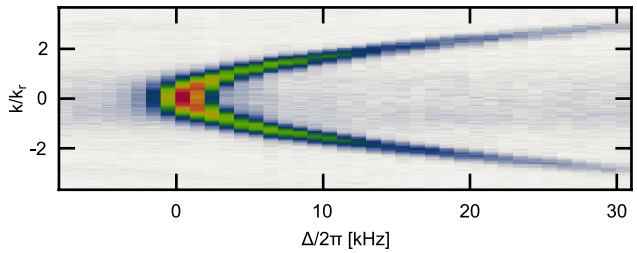
34. Pertot, D., Greif, D., Albert, S., Gadway, B. & Schneble, D. Versatile transporter apparatus for experiments with optically trapped Bose–Einstein condensates. *J. Phys. B* **42**, 215305 (2009).
35. Campbell, G. K. et al. Imaging the Mott insulator shells by using atomic clock shifts. *Science* **313**, 649–652 (2006).
36. Gadway, B., Pertot, D., Reimann, R., Cohen, M. G. & Schneble, D. Analysis of Kapitza–Dirac diffraction patterns beyond the Raman–Nath regime. *Opt. Express* **17**, 19173–19180 (2009).
37. Krinner, L., Stewart, M., Pazmiño, A. & Schneble, D. In situ magnetometry for experiments with atomic quantum gases. *Rev. Sci. Instrum.* **89**, 013108 (2018).
38. Deng, L. et al. Temporal, matter-wave-dispersion Talbot effect. *Phys. Rev. Lett.* **83**, 5407–5411 (1999).

© 2018 Springer Nature Limited. All rights reserved.



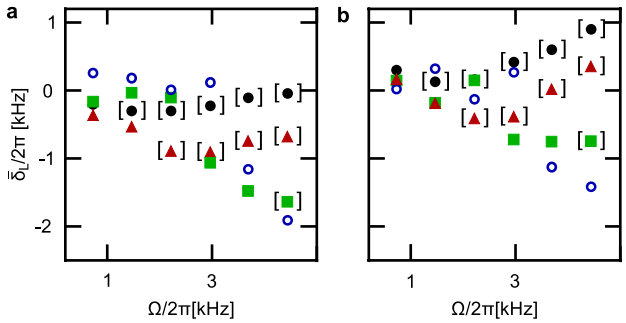
Extended Data Fig. 1 | Average spectrum of coupling stationary atoms into the *z* lattice. The spectrum was generated from a series of 17 spectra taken over a one-day period, whose fitted centres are shifted to zero. The coupling strength is $\Omega = 740(10)$ Hz and the pulse time is 400 μs, with the data points binned into 300-Hz-wide bins. The solid curve is a fit to the data with Ω as a free-fitting parameter and the dashed curve has no free parameters. The effective coupling strength was calculated using the wavefunction overlap between free and trapped species, $\gamma_0 = 0.72$.

© 2018 Springer Nature Limited. All rights reserved.



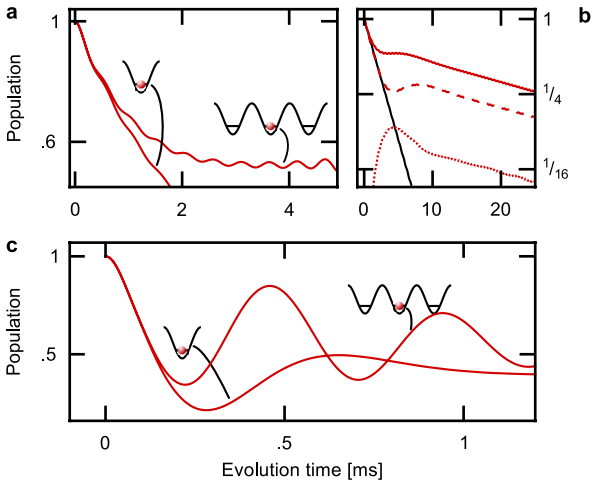
Extended Data Fig. 2 | Raw momentum spectrum. The spectrum shows a detuning-independent, diffuse background of roughly 10^3 atoms. The spectrum was acquired as described in the main text, Fig. 2 and Methods; colour scale is identical to Fig. 2c.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 3 | Raw data used to obtain the energy shift. **a**, **b**, Second moment of k (**a**) and half-separation squared (**b**) both subtracted by $\Delta/(2\pi)$. The detuning $\Delta/(2\pi)$ is 1.0 kHz (black disks), 2.0 kHz (red triangles), 4.0 kHz (green squares) and 6.0 kHz (blue circles). Points in brackets correspond to the non-Markovian regime, $\Omega/\Delta > 1$.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 4 | Simulated decay dynamics for a 1-site and a 3-site model (with the central site initially populated). a, Dynamics of the two models, as depicted in the insets, for $\Delta = 2\pi \times 1.9$ kHz and $\Omega = 2\pi \times 0.74$ kHz, with $\omega_z = 2\pi \times 0.1$ kHz. **b**, Long-time decay dynamics of the 1-site (black) and 3-site (red) models for $\Omega = 2\pi \times 0.74$ kHz and $\Delta = 2\pi \times 1.9$ kHz, with $\omega_z = 2\pi \times 5$ Hz. The dashed red line shows the population of the central, initially populated, site; the dotted red line shows the population of the neighbouring sites. **c**, Dynamics of the two models for $\Delta = -2\pi \times 0.1$ kHz and $\Omega = 2\pi \times 3$ kHz, with $\omega_z = 2\pi \times 0.1$ kHz.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0332-7

Complex silica composite nanomaterials templated with DNA origami

Xiaoguo Liu[1,2,10], Fei Zhang[3,4,10], Xinxin Jing[1,10], Muchen Pan[1], Pi Liu[5,6], Wei Li[7], Bowen Zhu[1], Jiang Li[1,8], Hong Chen[9], Lihua Wang[1], Jianping Lin[5,6], Yan Liu[3,4], Dongyuan Zhao[7], Hao Yan[3,4]* & Chunhai Fan[1]*

Genetically encoded protein scaffolds often serve as templates for the mineralization of biocomposite materials with complex yet highly controlled structural features that span from nanometres to the macroscopic scale[1–4]. Methods developed to mimic these fabrication capabilities can produce synthetic materials with well defined micro- and macro-sized features, but extending control to the nanoscale remains challenging[5,6]. DNA nanotechnology can deliver a wide range of customized nanoscale two- and three-dimensional assemblies with controlled sizes and shapes[7–11]. But although DNA has been used to modulate the morphology of inorganic materials[12,13] and DNA nanostructures have served as moulds[14,15] and templates[16,17], it remains challenging to exploit the potential of DNA nanostructures fully because they require high-ionic-strength solutions to maintain their structure, and this in turn gives rise to surface charging that suppresses the material deposition. Here we report that the Stöber method, widely used for producing silica (silicon dioxide) nanostructures, can be adjusted to overcome this difficulty: when synthesis conditions are such that mineral precursor molecules do not deposit directly but first form clusters, DNA–silica hybrid materials that faithfully replicate the complex geometric information of a wide range of different DNA origami scaffolds are readily obtained. We illustrate this approach using frame-like, curved and porous DNA nanostructures, with one-, two- and three-dimensional complex hierarchical architectures that range in size from 10 to 1,000 nanometres. We also show that after coating with an amorphous silica layer, the thickness of which can be tuned by adjusting the growth time, hybrid structures can be up to ten times tougher than the DNA template while maintaining flexibility. These findings establish our approach as a general method for creating biomimetic silica nanostructures.

We illustrate our strategy using a DNA origami silicification (DOS) nanostructure with a pattern similar to the cell wall unit of diatoms (Supplementary Figs. 1 and 2), with Fig. 1a sketching the silicification process. First, we designed and constructed a DNA origami structure containing pores with theoretical inner diameters of 5.4 nm, 9.8 nm and 82.4 nm, respectively (Supplementary Fig. 3). In contrast to the conditions under which previous double-stranded DNA (dsDNA)-guided silicification was performed[12,13], DNA origami nanostructures need a high concentration of cations such as Mg^{2+} to maintain their structural integrity[8]. But the high ion strength also prevents the attachment of *N*-trimethoxysilylpropyl-*N*,*N*,*N*-trimethylammonium chloride (TMAPS) onto the negatively charged phosphate backbone of the DNA that is needed for silicification (Supplementary Fig. 4).

Molecular dynamics simulation revealed that hydrolysed clusters T_{ab} (where *a* and *b* are the numbers of TMAPS and tetraethyl orthosilicate (TEOS) molecules, respectively) with more than three TMAPS can overcome the electrostatic potential barrier and adsorb to the DNA backbone under high-ionic-strength conditions (Supplementary Fig. 5 and Supplementary Table 1). We therefore adapted the synthesis method to include a pre-hydrolysis step that promotes the formation of T_{ab} clusters that are able to form an inductive layer around the DNA nanostructures, thus allowing for silica deposition. Optimized synthesis solutions had a TMAPS to TEOS stoichiometry set at 36.0 mM: 90.0 mM (Supplementary Figs. 6–9 and Supplementary Table 2), which ensures that pre-folded DNA nanostructures can provide sufficient surface area and enough negative charge density to recruit T_{ab} clusters before they self-aggregate with free TEOS in the bulk solution. This was confirmed by coarse-grained molecular dynamics simulations using a 198-base-pair (bp) three-dimensional DNA origami fragment along with a 348-bp two-dimensional DNA origami fragment that reacted with T_{ab} clusters (Fig. 1b and c, Supplementary Figs. 10–13 and Supplementary Table 1). Our finding that the initial formation of small clusters is an essential step in the silicification process is in accord with earlier observations, which concluded that amorphous silica nanoparticle synthesis involves colloidally stabilized primary particles that associate with and subsequently add to the silica structure[18].

Figure 2 gives examples of DNA origami nanostructures of differing complexity characterized by conventional transmission electron microscopy (TEM) and atomic force microscopy (AFM) before and after silicification. TEM imaging of simple DNA origami triangles in Fig. 2a reveals the successful formation of both the original DNA origami templates and the DOS nanostructures, with measured average edge lengths of the triangular DNA origami frameworks and DOSs of 120.2 ± 2.3 nm ($N = 20$) and 122.9 ± 4.1 nm ($N = 20$), respectively. The average DNA origami edge width was measured as 22.6 ± 1.5 nm ($N = 20$) while the designed edge width and measured DOS edge widths were 26.0 nm and 25.9 ± 1.4 nm, respectively, indicating edge shrinking for the stained pure DNA sample under TEM of around 3.4 nm. Such shrinking is often seen when DNA origami is dried in air on diverse substrates[9,16,19] (see Supplementary Fig. 14, which also documents the height changes and dependence of changes on imaging mode and conditions). The heights and side lengths of the triangular DOS, based on AFM and TEM images, show the formation of a homogeneous thin silica coating layer with uniform thickness (mean value 3.1 nm for the *z* axis and 2.7 nm for the *x*–*y* plane, Fig. 2b and Supplementary Figs. 15–18; see also Supplementary Fig. 19, with additional discussion on surface roughness). The sample composition was analysed by energy dispersive spectrometer (EDS) mapping, with an overlay of the obtained elemental mapping images with scanning electron microscope (SEM) images further confirming the uniform deposition of silica on the triangular DNA framework (Fig. 2c).

[1]Division of Physical Biology and Bioimaging Center, Shanghai Synchrotron Radiation Facility, CAS Key Laboratory of Interfacial Physics and Technology, Shanghai Institute of Applied Physics, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shanghai, China. [2]School of Chemistry and Chemical Engineering, and Institute of Molecular Medicine, Renji Hospital, School of Medicine, Shanghai Jiao Tong University, Shanghai, China. [3]Center for Molecular Design and Biomimetics, Biodesign Institute, Tempe, AZ, USA. [4]School of Molecular Sciences, Arizona State University, Tempe, AZ, USA. [5]State Key Laboratory of Medicinal Chemical Biology, College of Pharmacy and Tianjin Key Laboratory of Molecular Drug Research, Nankai University, Haihe Education Park, Tianjin, China. [6]Biodesign Center, Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin, China. [7]Department of Chemistry, Laboratory of Advanced Materials, Shanghai Key Lab of Molecular Catalysis and Innovative Materials, iChEM and State Key Laboratory of Molecular Engineering of Polymers, Fudan University, Shanghai, China. [8]Shanghai Key Laboratory of Green Chemistry and Chemical Processes, School of Chemistry and Molecular Engineering, East China Normal University, Shanghai, China. [9]Pen-Tung Sah Institute of Micro-Nano Science and Technology, Xiamen University, Xiamen, China. [10]These authors contributed equally: Xiaoguo Liu, Fei Zhang, Xinxin Jing. *e-mail: hao.yan@asu.edu; fchh@sinap.ac.cn

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

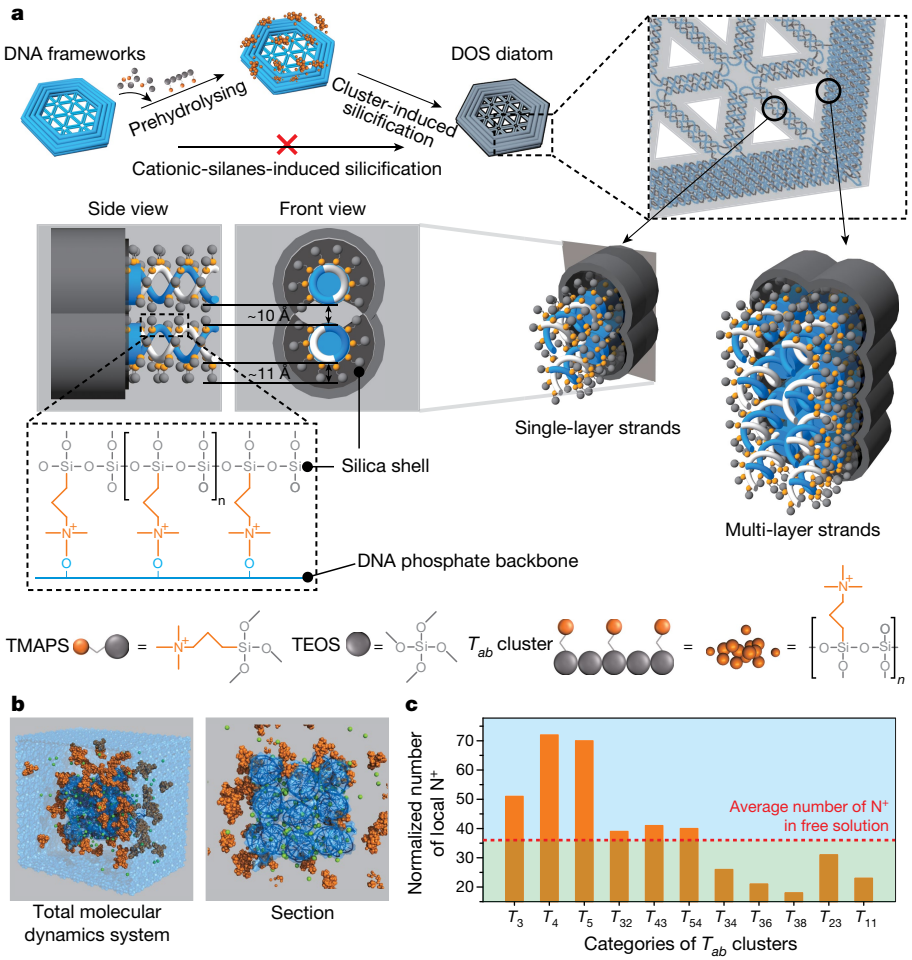


Fig. 1 | Schematic illustration and molecular dynamics simulation of DOS strategy. a, The reaction process of the DOS diatom-mimicking structure. The silicification was achieved through mixing DNA framework with prehydrolysed TMAPS and TEOS. The prehydrolysing cluster has multiple TMAPS molecules (noted as T_{ab}, where a, $b = 0, 1, 2, 3 \ldots$). The configurations of the individual DNA double helices in the complex DNA origami templates could be identified as single- and multi-layered strands. In both cases, the silica could not be inserted into the spaces between the DNA helices, because the precursor molecule was much bigger than the gaps (about 5–10 Å) between the two adjacent dsDNA molecules. Also, there was a high Mg^{2+} density inside DNA nanostructures. **b**, Molecular dynamics simulation shows the DOS process whereby T_{ab} clusters (for example, T_4) adhere onto the DNA nanostructure surfaces. A 198-bp three-dimensional DNA origami fragment was used as our model. Mg^{2+} (green), DNA strands (blue), water box (light blue) and T_{ab} clusters (orange) are shown. **c**, The simulation revealed that a minimal amount of T_3 precursor ($a \geq 3$) is required to compete with the high concentration of cations (for example, 12.5 mM Mg^{2+}) that was required by the DNA nanostructures. However, over-hydrolysed clusters such as T_{38} tend to self-aggregate before being recruited by the DNA origami. The normalized number of local nitrogen ions (that is, equal in concentration to the TMAPS, within an 8 Å range of the DNA origami surface), is higher than the average number of nitrogen ions (36) found in the free solution that contains the DNA origami, which indicates the successful attachment of the clusters onto the DNA backbones.

To demonstrate that our method retains the ability of DNA architecture-encoded nanofabrication for controlling sample morphology with nanometre resolution[14,15], we built a set of DNA origami squares containing nanopores with sizes ranging, theoretically, from 31 nm × 44 nm to 25 nm × 33 nm and 13 nm × 11 nm (not taking the un-hybridized M13 fragment and flexibility of the DNA nanostructure into consideration; see Supplementary Figs. 20–22 for further design details). As illustrated in Fig. 2d, measured origami pore sizes were 639.9 nm^2, 283.8 nm^2 and 103.6 nm^2 from class-averaged TEM images (Supplementary Figs. 23 and 24). After silicification, pore sizes changed to 608.8 nm^2, 255.7 nm^2 and 87.5 nm^2 as determined from representative averaged TEM images ($N = 39$, Supplementary Fig. 25); these values correspond to diameters of 27.8 nm, 18.4 nm and 10.6 nm for a circular pore, and to side lengths of 24.7 nm, 16.0 nm and 9.4 nm for a square opening.

Our method can in principle create single nanopores down to 3 × 3 nm^2, which is based on the diameter of the original DNA helices and the thickness of the silica layer after growth. This is in fact illustrated by the diatom nanostructure, finally shown in Fig. 2e (see also Supplementary Fig. 26), with the negatively strained DNA nanostructure before silica deposition seen to have a less defined morphology and twisted outer edges. In contrast, the class-averaged TEM image ($N = 20$) reveals well replicated outer edges and two sets of inner pores with diameters of about 5.5 nm and 3.4 nm that agree well with the design (more details in Supplementary Figs. 27 and 28 and Supplementary Tables 3–5). The area of the smaller pore, about 12.0 nm^2, is much smaller than the approximately 47.8 nm^2 previously reported for the smallest nanopore obtained by the DNA origami lithography method[17] and represents a resolution comparable to that achieved with electron beam methods[20].

Beyond accurately controlling the final silica shell thickness and pore size, our strategy provides a general and robust procedure for building rigid three-dimensional inorganic structures in arbitrary shapes that are difficult to create using lithography or polymer coating[14,15,17,21]. This is illustrated in Fig. 3a, which shows the basic designs and negatively stained DNA architectures (see also Supplementary Figs. 29–31) and Fig. 3b and c, which shows the corresponding DOS structures that include two-dimensional units (square, triangle, cross and diatom-mimicking), the two-dimensional honeycomb pattern and

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

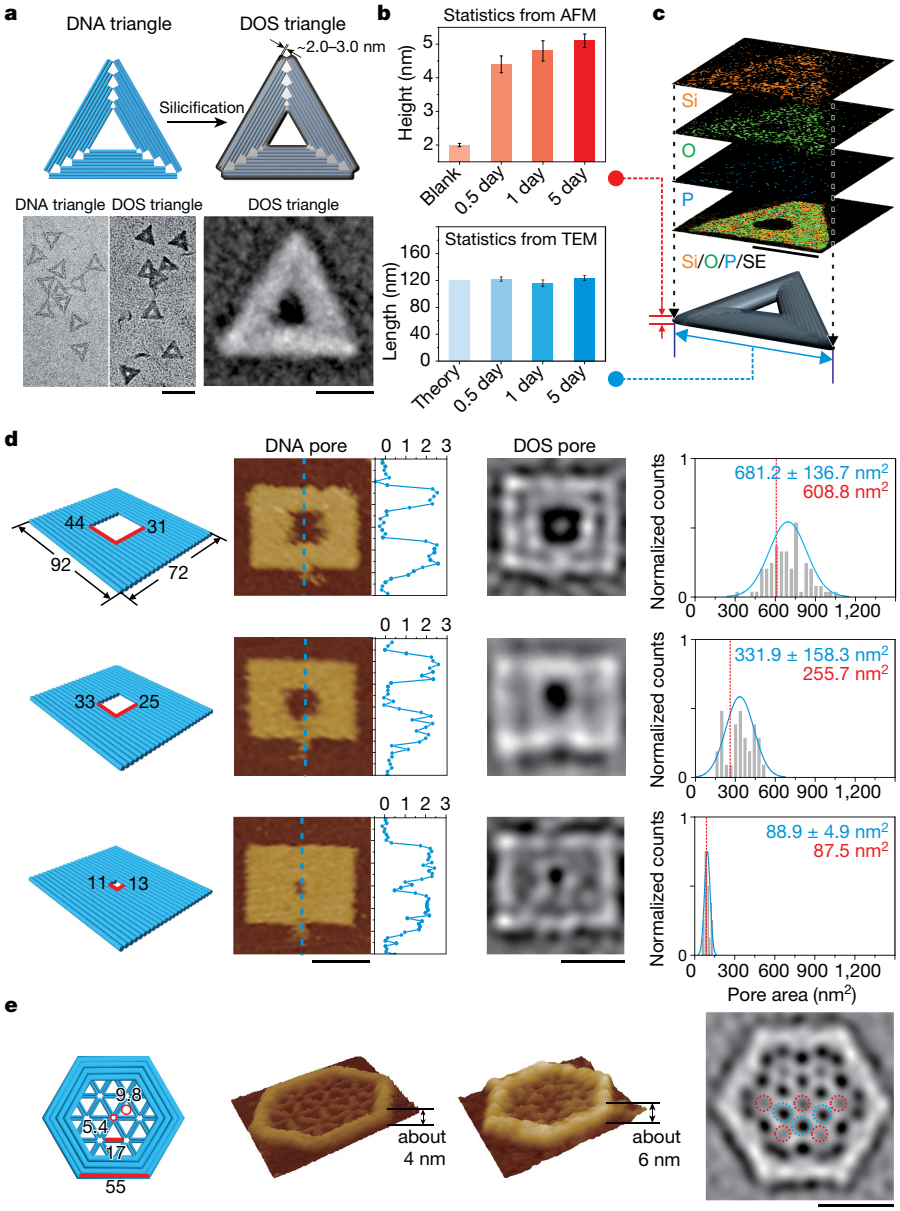


Fig. 2 | Geometrically precise control of DOS structures. a, A triangular DOS nanostructure as a representative structure showed near-perfect shape replication from DNA origami templates to DOS nanostructures (schematic at top); the bottom left panel shows the negatively stained TEM image of the DNA origami templates; the bottom middle panel shows the TEM image of the DOS triangle; and the bottom right panel is the class-averaged TEM image of the DOS triangle. Scale bars, zoomed-out 200 nm and zoomed-in 50 nm. **b**, Statistics of the heights (from AFM) and lengths (from TEM) for DOS nanostructures with increasing growth time. Error bars, s.d. **c**, EDS mapping of the DNA–silica hybrid nanostructure (Si is coloured in orange, O in green and P in blue; SE, secondary-electron image). Scale bar, 50 nm. **d**, A user-specified DOS nanopore with the smallest diameter down to less than 10 nm. From left to right, panels show the design, AFM images for pure DNA origami, the class-averaged TEM images, and the size distribution of silica nanopore measured from raw TEM images, respectively. Histograms of three different pore sizes were normalized and fitted to Gaussian distribution curves (light blue), and the red dashed lines indicate direct measurements of the pore sizes from class-averaged TEM images. **e**, DOS diatom nanostructures. From left to right, panels show the design model of a DNA origami template, AFM images before and after silicification, and the averaged TEM image of a DOS diatom nanostructure, respectively. Scale bars, 50 nm. All unmarked units refer to nanometres.

three-dimensional objects (cube, tetrahedron, hemisphere, toroid and ellipsoid). The parameters of each DOS structure measured from TEM images were compared with their original DNA framework design in Supplementary Table 6. See Supplementary Figs. 32–56, Supplementary Videos 1–3 and additional discussion for details on characterization, comparison to previous methods, growth kinetics and yields.

Although the mechanical properties of DNA nanostructures are important in DNA-assisted assembly and nanofabrication[7,22] and the tensile strength of DNA has been probed[23,24], little is known about the mechanical properties of DNA under pressure[25,26]. We found, using AFM Peak Force toolkit (see Supplementary Information, Methods) measurements, a tenfold increase of the Young's modulus (*E* modulus) from around 100 MPa for a DNA origami to around 1 GPa for the corresponding two-dimensional DOS triangle, and a greatly enhanced toughness that could withstand setoff forces up to 3,000 pN (Fig. 4a and b, Supplementary Figs. 57–64). Although the DOS modulus is below the values of 10–100 GPa reported for nanoporous silica[27], there might be scope for improving it. Figure 4c details the mechanical response of a three-dimensional DOS tetrahedron under compression, which was measured by Peak Force quantitative nanomechanical mapping in fluid mode and revealed an increase in the average peak heights of the samples from 16.5 nm, 22.9 nm and 39.7 nm to 52.1 nm ($N = 10$ nN

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

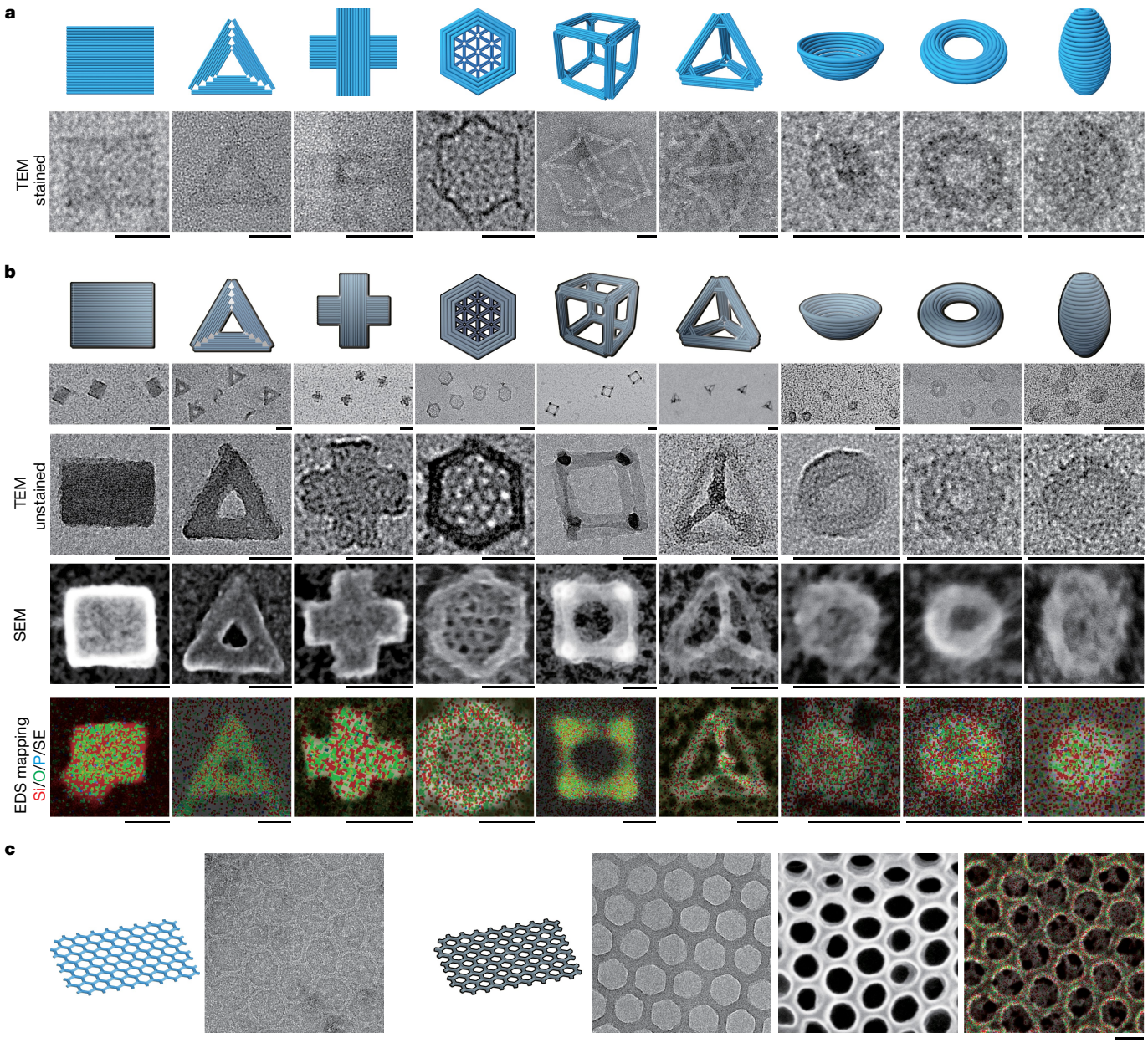


Fig. 3 | Representative complex DOS nanostructures. a, Designed models and corresponding TEM images of the DNA origami templates. From left to right, panels show the two-dimensional structures (square, triangle, cross and DOS diatom) and three-dimensional objects (cube, tetrahedron, hemisphere, toroid and ellipsoid). **b**, The corresponding DOS nanostructures. From top to bottom, panels show TEM images (zoomed-out and zoomed-in), SEM images and EDS mapping images (Si is coloured in red, O in green and P in blue; SE, secondary-electron image). **c**, Silicification of a honeycomb two-dimensional pattern. Left: negatively stained TEM image of DNA-only assemblies; right: TEM, SEM and EDS images of the DOS honeycomb. Scale bars, zoomed-out 100 nm and zoomed-in 50 nm.

and 1.0 nN) after 0-, 1-, 2- and 5-day silica growth, respectively (Supplementary Figs. 65–67). Although the maximum measured height was below the theoretical height for a rigid ideal tetrahedron (74.0 nm), the results nonetheless illustrate that pure DNA origami nanostructures placed on a mica surface deform dramatically and that DOS nanostructures have enhanced rigidity that increases with silica growth time.

With the dsDNA tetrahedron reported to maintain its structural integrity[26] up to a maximum compression force of about 100 pN, we next studied the load-deformation behaviour of a tetrahedron DOS and found no obvious shape changes for forces up to 1,000 pN (Supplementary Fig. 68). Even a 3.0 nN load force only bent the structure without damaging it. As shown in Fig. 4d and Supplementary Fig. 69, the larger applied tip forces can lead to greater bending of the DOS tetrahedrons that had experienced 5 days of silica growth.

As force-deformation responses showed a nearly ideal elastic behaviour, we further explored the elastic properties of the DOS tetrahedrons by applying force jumping cycles between 3.0 nN and 1.0 nN (Fig. 4e) and found that DOS tetrahedrons withstood at least 5 cycles and restored 80% of their initial height before being irreversibly damaged (Supplementary Fig. 70). The DOS nanostructures thus possess toughness and structural flexibility, that is, mechanical properties typically found in biominerals[28]. We attribute the enhanced rigidity of the DOS nanostructures (see also Supplementary Figs. 71–77) to the amorphous silica shell, with further structural support provided by water when in a liquid environment and by the bonds between the DNA molecule and the silicon–oxygen tetrahedron of the inorganic phase. In a final exploration of silicification-enhanced rigidity, we attached gold nanorods to two edges of the tetrahedra and found that the free-standing DOS–Au nanorods structure retains a distortion-free

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

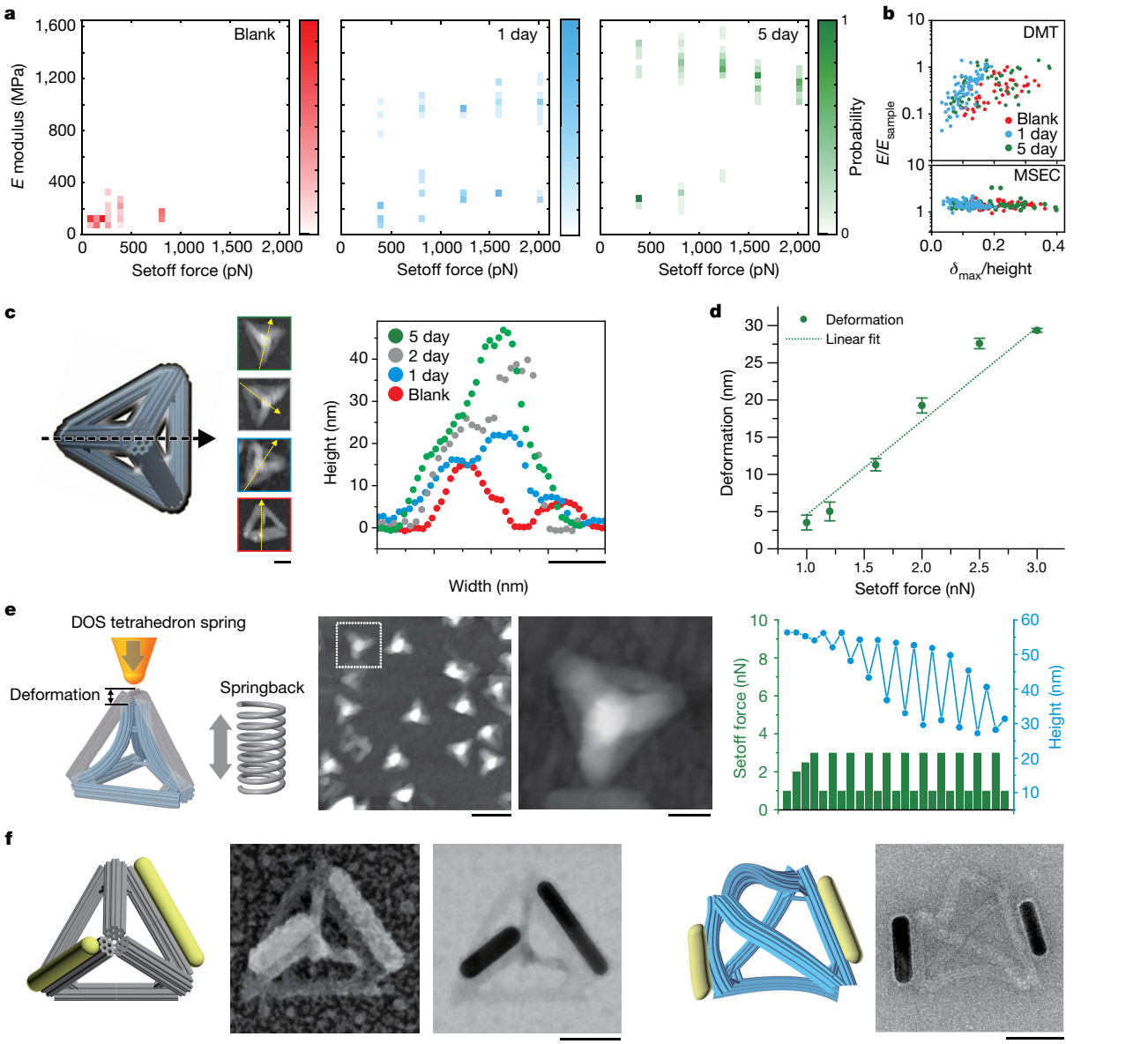


Fig. 4 | Nanomechanical studies on DOS nanostructures. a, Young's modulus (*E* modulus) derived from force curves, using a membrane substrate effect correction (MSEC) model. The *E* modulus of a DOS triangle after 5 days' silicification was about ten times greater than the original DNA nanostructure. **b**, Comparison between a Derjaguin–Muller–Toporov (DMT) model and an MSEC model. 'δ_{max}/height' indicates that the maximum indentation depth was divided by the sample thickness; E/E_{sample} indicates the dispersion of the *E* modulus when taking substrate effects into consideration. The *E* modulus data that were deducted from the MSEC model were independent of thickness and force. **c**, AFM-tip-induced deformation of a DOS tetrahedron. The pure DNA tetrahedron was deposited on the mica substrate, with a maximum height of 15.2 nm. Samples aged for 1, 2 and 5 days showed increasing heights of 21.9 nm, 44.1 nm and 48.5 nm correspondingly and were able to stand freely on the mica surface. The AFM tip force used for the measurement was 1.0 nN. The arrows indicate the height section directions. Scale bars, 50 nm. **d**, Height changes during the deformation of a single DOS tetrahedron that was grown for 5 days, under increasing tip forces, showed changes from 1.0 nN to 3.0 nN. Error bars, s.d. **e**, Left, schematics of the DOS tetrahedron spring. Centre, AFM images (tapping in fluid) of DOS tetrahedron after 5 days' silicification. Cycling forces were applied to a single DOS tetrahedron and the recovery of structures further confirmed the unique elastic mechanical property of the DOS nanostructures. Scale bars, zoomed-out 200 nm and zoomed-in 50 nm. **f**, A free-standing DOS–Au nanorod tetrahedron that faithfully replicates the original design, as compared to a DNA origami tetrahedron supported nanorod structure, which collapsed onto the surface owing to its flexibility. Left, SEM image; centre, STEM image; and far right, TEM image. Scale bars, 50 nm.

tetrahedral geometry while the DNA-origami–Au nanorods structure collapsed (Fig. 4f, Supplementary Figs. 78–80, Supplementary Video 4). This suggests that DOS frameworks could in principle support metal nanoparticles in three-dimensional space with defined angles and coupling distances that might be of interest for developing nano-electronic and nano-photonic devices, with distortion-free nanoscale geometries much better than have been achieved in previously reported DNA–metal nanostructures[7].

In closing, we note that one long-standing aim of nanotechnology is full control over structural features with nanometre or even atomic-level precision. DNA has been much studied in that context as it can be readily programmed to assemble into well defined customized nanostructures of considerable complexity. The method we have presented here offers the possibility of transferring nanoscale geometric information from designer DNA nanostructures (especially framework nucleic acids) to inorganic materials, illustrated by the successful formation of silica–DNA composites with a wide range of different shapes and inorganic coating thicknesses that control the rigidity of the structures. Provided suitable synthesis conditions can be found, it should in principle be possible to create composites where other

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

inorganic materials coat DNA origami shapes; or even multi-component composites, through stepwise deposition of more than one inorganic material on DNA templates[9,14,29,30].

Data availability

All data generated or analysed during this study are included in the paper and its Supplementary Information, and are available from the corresponding author on request.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0332-7.

Received: 1 November 2017; Accepted: 14 May 2018;
Published online 16 July 2018.

1. Berman, A. et al. Biological control of crystal texture: a widespread strategy for adapting crystal properties to function. *Science* **259**, 776–779 (1993).
2. Belcher, A. M. et al. Control of crystal phase switching and orientation by soluble mollusc-shell proteins. *Nature* **381**, 56–58 (1996).
3. Aizenberg, J. et al. Skeleton of *Euplectella* sp.: structural hierarchy from the nanoscale to the macroscale. *Science* **309**, 275–278 (2005).
4. Kröger, N., Deutzmann, R. & Sumper, M. Polycationic peptides from diatom biosilica that direct silica nanosphere formation. *Science* **286**, 1129–1132 (1999).
5. Sumper, M. Biomimetic patterning of silica by long-chain polyamines. *Angew. Chem. Int. Ed.* **43**, 2251–2254 (2004).
6. Wegst, U. G., Bai, H., Saiz, E., Tomsia, A. P. & Ritchie, R. O. Bioinspired structural materials. *Nat. Mater.* **14**, 23–36 (2015).
7. Jones, M. R., Seeman, N. C. & Mirkin, C. A. Programmable materials and the nature of the DNA bond. *Science* **347**, 1260901 (2015).
8. Rothemund, P. W. K. Folding DNA to create nanoscale shapes and patterns. *Nature* **440**, 297–302 (2006).
9. Kershner, R. J. et al. Placement and orientation of individual DNA shapes on lithographically patterned surfaces. *Nat. Nanotechnol.* **4**, 557–561 (2009).
10. Zheng, J. et al. From molecular to macroscopic via the rational design of a self-assembled 3D DNA crystal. *Nature* **461**, 74–77 (2009).
11. Yang, F., Zuo, X., Fan, C. & Zhang, X. Biomacromolecular nanostructures-based interfacial engineering: from precise assembly to precision biosensing. *Nat. Sci. Rev.* nwx134 https://doi.org/10.1093/nsr/nwx134 (2018).
12. Numata, M., Sugiyasu, K., Hasegawa, T. & Shinkai, S. Sol-gel reaction using DNA as a template: an attempt toward transcription of DNA into inorganic materials. *Angew. Chem. Int. Ed.* **43**, 3279–3283 (2004).
13. Jin, C., Qiu, H., Han, L., Shu, M. & Che, S. DNA transcription into diverse porous silicas by a co-structure directing route: chiral, ring and ordered nanochannel arrays. *Chem. Commun*. 3407–3409 (2009).
14. Peng, Z. & Liu, H. Bottom-up nanofabrication using DNA nanostructures. *Chem. Mater.* **28**, 1012–1021 (2016).
15. Sun, W. et al. Casting inorganic structures with DNA molds. *Science* **346**, 1258361 (2014).
16. Surwade, S. P. et al. Nanoscale growth and patterning of inorganic oxides using DNA nanostructure templates. *J. Am. Chem. Soc.* **135**, 6778–6781 (2013).
17. Diagne, C. T., Brun, C., Gasparutto, D., Baillin, X. & Tiron, R. DNA origami mask for sub-ten-nanometer lithography. *ACS Nano* **10**, 6458–6463 (2016).
18. Carcouët, C. C. et al. Nucleation and growth of monodisperse silica nanoparticles. *Nano Lett.* **14**, 1433–1438 (2014).
19. Kim, H., Surwade, S. P., Powell, A., O'Donnell, C. & Liu, H. Stability of DNA origami nanostructure under diverse chemical environments. *Chem. Mater.* **26**, 5265–5273 (2014).
20. Li, J. et al. Ion-beam sculpting at nanometre length scales. *Nature* **412**, 166–169 (2001).
21. Agarwal, N. P., Matthies, M., Gur, F. N., Osada, K. & Schmidt, T. L. Block copolymer micellization as a protection strategy for DNA origami. *Angew. Chem. Int. Ed.* **56**, 5460–5464 (2017).
22. Zhang, F., Nangreave, J., Liu, Y. & Yan, H. Structural DNA nanotechnology: state of the art and future perspective. *J. Am. Chem. Soc.* **136**, 11198–11211 (2014).
23. Bustamante, C., Bryant, Z. & Smith, S. B. Ten years of tension: single-molecule DNA mechanics. *Nature* **421**, 423–427 (2003).
24. Smith, S., Finzi, L. & Bustamante, C. Direct mechanical measurements of the elasticity of single DNA molecules by using magnetic beads. *Science* **258**, 1122–1126 (1992).
25. Zhou, X. F. et al. Radial compression elasticity of single DNA molecules studied by vibrating scanning polarization force microscopy. *Phys. Rev. E* **71**, 062901 (2005).
26. Goodman, R. P. et al. Rapid chiral assembly of rigid DNA building blocks for molecular nanofabrication. *Science* **310**, 1661–1665 (2005).
27. Fan, H. et al. Modulus-density scaling behaviour and framework architecture of nanoporous self-assembled silicas. *Nat. Mater.* **6**, 418–423 (2007).
28. Meyers, M. A., McKittrick, J. & Chen, P. Y. Structural biological materials: critical mechanics-materials connections. *Science* **339**, 773–779 (2013).
29. Wang, T. et al. A DNA crystal designed to contain two molecules per asymmetric unit. *J. Am. Chem. Soc.* **132**, 15471–15473 (2010).
30. Sha, R. et al. Self-assembled DNA crystals: the impact on resolution of 5′-phosphates and the DNA source. *Nano Lett.* **13**, 793–797 (2013).

Acknowledgements This Letter is dedicated to the memory of Q. Huang, who initiated the project along with C.F. and X.L. We thank Z. Liu (Shanghai Technical Platform for Testing and Characterization on Inorganic Materials, CASSIC, China) for providing technical support, Y. Li (South China Normal University, China) for providing diatom specimens, and the BL16B1/BL19U2 beamline at Shanghai Synchrotron Radiation Facility (SSRF) for ongoing support in data collection and molecular dynamics simulation. This project was supported by National Science Foundation of China (grant numbers 21390414, 21329501, 21603262 and 21675167), National Key R&D Program of China (grant numbers 2016YFA0201200 and 2016YFA0400900) and the Key Research Program of Frontier Sciences, CAS (grant number QYZDJ-SSW-SLH031). L.W., C.F. and H.Y. thank the National Key R&D Program of China (grant number 2016YFA0400900). H.Y., F.Z. and Y.L. thank the US National Science Foundation, Office of Naval Research, Army Research Office, National Institutes of Health, and Department of Energy for financial support. W.L. and D.Z. thank the National Science Foundation of China (grant numbers U1463206 and 21733003), National Key R&D Program of China (grant number 2018YFA0209401) and Key Basic Research Program of the Science and Technology Commission of Shanghai Municipality (grant number 17JC1400100).

Author contributions C.F. and H.Y. supervised the research. X.L., C.F. and H.Y. conceived the research, designed the experiments. F.Z. designed the DNA nanostructures. X.J. and X.L. carried out silicification experiments and characterization. M.P. analysed EM and AFM data. P.L. carried out molecular dynamics simulations. W.L. and D.Z. provided silicification characterization and discussion. B.Z. carried out FTIR and Raman experiments. All authors analysed data. X.L., F.Z., X.J., M.P., D.Z., C.F. and H.Y. interpreted data and wrote the paper.

Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0332-7.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to H.Y. and C.F.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0328-3

Atmosphere–soil carbon transfer as a function of soil depth

Jérôme Balesdent[1]*, Isabelle Basile-Doelsch[1], Joël Chadoeuf[2], Sophie Cornu[1], Delphine Derrien[3], Zuzana Fekiacova[1] & Christine Hatté[4]

The exchange of carbon between soil organic carbon (SOC) and the atmosphere affects the climate[1,2] and—because of the importance of organic matter to soil fertility—agricultural productivity[3]. The dynamics of topsoil carbon has been relatively well quantified[4], but half of the soil carbon is located in deeper soil layers (below 30 centimetres)[5–7], and many questions remain regarding the exchange of this deep carbon with the atmosphere[8]. This knowledge gap restricts soil carbon management policies and limits global carbon models[1,9,10]. Here we quantify the recent incorporation of atmosphere-derived carbon atoms into whole-soil profiles, through a meta-analysis of changes in stable carbon isotope signatures at 112 grassland, forest and cropland sites, across different climatic zones, from 1965 to 2015. We find, in agreement with previous work[5,6], that soil at a depth of 30–100 centimetres beneath the surface (the subsoil) contains on average 47 per cent of the topmost metre's SOC stocks. However, we show that this subsoil accounts for just 19 per cent of the SOC that has been recently incorporated (within the past 50 years) into the topmost metre. Globally, the median depth of recent carbon incorporation into mineral soil is 10 centimetres. Variations in the relative allocation of carbon to deep soil layers are better explained by the aridity index than by mean annual temperature. Land use for crops reduces the incorporation of carbon into the soil surface layer, but not into deeper layers. Our results suggest that SOC dynamics and its responses to climatic control or land use are strongly dependent on soil depth. We propose that using multilayer soil modules in global carbon models, tested with our data, could help to improve our understanding of soil–atmosphere carbon exchange.

The size of the Earth SOC reservoir is estimated to be around 1,500 gigatonnes of carbon (Gt C) in the first metre, excluding permafrost areas[6], making it a huge potential source or sink for atmospheric carbon (levels of which are increasing by +4.4 Gt C per year)[11]. The future response of this soil compartment could substantially affect not only the climate but also global food production (through the role of organic matter in soil fertility), as well as the stability or resilience of ecosystems[3]. About half of this carbon is located at depths below 30 cm (refs [5,6]). However, although the dynamics of topsoil carbon has been relatively well quantified, especially thanks to long-term experiments carried out over generations[4], major questions remain about how to estimate changes in deep-soil carbon and the processes involved. Decision-makers and ecosystem managers are thus deprived of any references for the management of the deep carbon stock. Similarly, when modelling the Earth system and the global carbon cycle, the scientific community also constantly faces the problem of modelling the dynamics of deep carbon[9,10,12].

Neither absolute changes in carbon stocks nor carbon fluxes in the deep soil horizons can be quantified by direct measurement. Owing to the very low carbon concentrations (on average less than 5 g kg^{-1} at depths of 80 cm), spatial heterogeneity and slow changes, temporal variations in stocks are smaller than measurement accuracy. Evidence for deep carbon changes is therefore exceptional[13,14]. Information on incoming fluxes resulting from root mortality and exudation by living roots is not accessible without tracers. In addition, in situ quantification of the outflow from the organic reservoir—which occurs mainly through heterotrophic respiration by organic-matter decomposers, in the form of CO_2 production—is very difficult, if not impossible, because the CO_2 efflux blends heterotrophic respiration and root autotrophic respiration[15]. Isotopic methods are therefore appropriate for tracing deep carbon dynamics. The radiocarbon age of deep carbon is indicative of its slow turnover[16–19], but ^{14}C dating, which provides mean ages, does not estimate the exact proportions of active and stable carbon[17,20]. Here we propose a stable-isotope-based observation of the actual depth distribution of soil carbon ages. It relies on sites that are marked by a natural change in the $^{13}C/^{12}C$ ratio of the vegetation at a known date. This is equivalent to the continuous in situ labelling of the atmospheric carbon atoms that have been incorporated into soil organic matter for a known duration, have eventually replaced pre-existing organic carbon, and have been retrieved at the date of sampling[21].

We conducted a meta-analysis of 112 such sites (Extended Data Fig. 1), where the labelling ranged from 4 to 4,000 years. At each site, the technique provides an indication of carbon age—that is, the proportion of carbon that is younger than the labelling duration; meta-analysis of similar sites with varied durations provides an age probability distribution[17]. Our study includes most of the world's biomes except boreal zones, and is evenly distributed among forests, grasslands and croplands.

We quantified carbon distribution in the two-dimensional age–depth continuum[22], the depth distribution of carbon incorporation in soil over the past 50 years, and the dependence of these factors on climate and land use. We also summarized depth distributions in terms of two layers, 0–30 cm (topsoil) and 30–100 cm (subsoil)—an arbitrary cut-off, but one that is often used in carbon inventories[6]. Our results, which are based on original observations, are independent of any datasets or modelling results from other studies.

Figure 1 depicts individual data showing the proportion of new carbon—that is, the proportion of SOC that derives from new vegetation—as a function of time. At all depths, a minor proportion of soil carbon is renewed rapidly (within ten years). Nine sites at which a vegetation signature change occurred more than 1,000 years ago reveal the incomplete replacement of carbon, that is, the presence of millennia-old soil carbon, at depth but not in the topsoil.

The rate of carbon incorporation in the topsoil was, as expected, strongly dependent on environmental variables, in particular land use ($P < 0.001$) and mean annual temperature (MAT; $P < 0.05$) (Extended Data Table 1). For the subsoil, by contrast, we found no relationship between carbon age and land use, and only a weak relationship with temperature ($P = 0.1$); instead, carbon age was more affected by the ratio of precipitation to potential evapotranspiration[23] ($P < 0.01$; Extended Data Table 2). This observation reinforces the results of ref. [9], which

[1]Aix-Marseille Université, CNRS, IRD, INRA, Coll France, CEREGE, Aix en Provence, France. [2]INRA UR 1052, Avignon, France. [3]INRA UR Biogéochimie des Ecosystèmes Forestiers, Nancy, France. [4]Laboratoire des Sciences du Climat et de l'Environnement, UMR 8212 CEA-CNRS-UVSQ, Université Paris-Saclay, Gif-sur-Yvette, France. *e-mail: jerome.balesdent@inra.fr

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

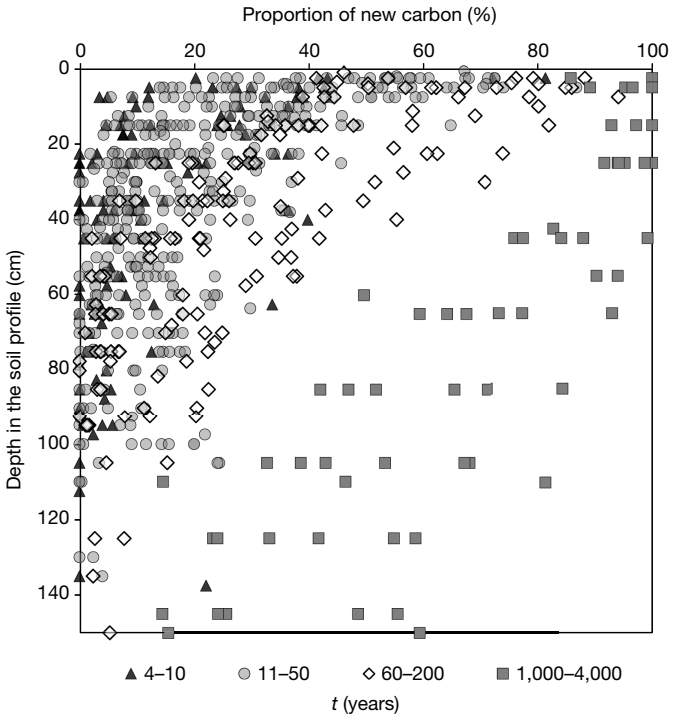


Fig. 1 | Observed proportions of new carbon in 112 soil profiles. For each soil sample, the proportion p of new carbon atoms was determined by the change in the soil carbon ^{13}C signature following a change in the ^{13}C signature of the vegetation for a given duration t; p is the proportion of carbon atoms younger than t[21]. Data are presented in four classes of duration t.

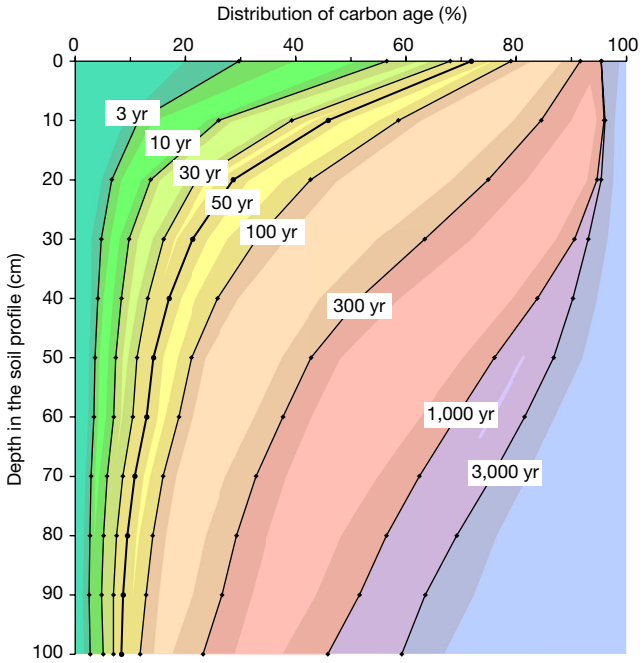


Fig. 2 | Meta-analysis of carbon age distribution over 55 tropical grassland and forest soil profiles. At each depth, the proportion of carbon aged less than time t (3 years, 10 years and so on) was fitted by a bi-exponential regression of t (Extended Data Table 3). Grey bands represent ±1 standard error of the estimated mean. The median age of soil carbon increases from seven years at depth 0 cm to 1,250 years at 100 cm. Integration of the carbon content in each layer demonstrates that the carbon of age less than 50 years represents 45% of topsoil carbon (0–30 cm) and 13% of deep carbon (30–100 cm).

showed that the relationship between ecosystem carbon turnover time and precipitation is pervasive and underestimated by models.

To analyse the age distribution with depth under comparable environmental conditions, we selected a homogeneous subset of sites, namely a group of forests and grasslands under warm and moist climates (with MATs higher than 17 °C, annual precipitation of more than 1,000 mm, and precipitation/evapotranspiration ratios greater than 0.8). Figure 2 and Extended Data Table 3 depict the detailed depth distribution of carbon ages throughout this panel of soils. This description of carbon dynamics in time–depth space highlights its strong dependence on both variables. The dynamics of subsoil carbon is around seven times slower than that of topsoil carbon (that is, it takes seven times longer to reach the same proportion of renewed carbon; Extended Data Fig. 2). In deep layers, the age distribution reveals the small but non-negligible direct incorporation of photosynthetically fixed carbon through deep roots or soluble carbon (for the youngest carbon), and the predominance of carbon that is older than 1,000 years. Mid-profile horizons (20–70 cm) are dominated by carbon of intermediate ages (100 to 1,000 years), which can be considered to result from the slow downward movement of carbon[16,24]. Carbon incorporation in the 100–200-cm layer has been quantified in only a few studies and averaged 5 ± 3% (1 standard deviation) of soil carbon after 50 years.

We calculated the amount of carbon incorporated into each layer (C_{new}, in units of kg C m^{-2}) for each site. In our database, the SOC found in the subsoil layer represents 47% of the total stock found in the entire top metre of soil, in agreement with the percentage of 47% to 52% reported globally[5,6]. To express the incorporation of new carbon in depth on the basis of a single indicator, we chose the ratio $R_{30\text{-}100}$, which is C_{new}(30–100 cm)/C_{new}(0–100 cm), and analysed its dependence on land use, climate and time in the 0- to 200-year-old sites (Extended Data Table 4). We found that $R_{30\text{-}100}$ is strongly dependent on land use ($P < 0.001$). The mean values of $R_{30\text{-}100}$ (50 years) are 19%, 21% and 29% for forests, grasslands and croplands, respectively. The relatively deeper carbon incorporation in croplands concerns all layers below a depth of 10 cm and cannot be explained only by soil mixing due to ploughing, as the depth of this mixing does not exceed 30 cm (Fig. 3). Croplands incorporate less new carbon in their topsoils than do forests and grasslands, whereas in subsoil the amount of incorporated carbon is similar (Extended Data Tables 1 and 2). This is consistent with the general reduction in carbon input at the soil surface[25] that results from the removal of above-ground biomass during harvesting. $R_{30\text{-}100}$ also depends on the precipitation/evapotranspiration index ($P < 0.005$), and is weakly dependent on MAT ($P < 0.1$; Extended Data Table 4), in accordance with the deeper rooting that takes place under dry climates, and possibly the more frequent occurrence of deep soils at low latitudes. The world average value of $R_{30\text{-}100}$ (50 years) is 19% (±4%; 95% confidence interval) (Fig. 3). The overall shallow incorporation of carbon can be expressed by the median depth of carbon incorporated in the last 50 years: 9 ± 1 cm (95% confidence interval) in forests, 10 ± 2 cm in grasslands and 17.5 ± 1.5 cm in croplands in our panel (9.7 ± 1.2 cm on average globally; Extended Data Table 5). Taking into account the 100–200-cm layer (when observed) would lower this median depth by 0.5 cm.

This study provides an unprecedented estimate of, first, the SOC age distribution over the soil profile (Fig. 2), and second, the depth distribution of the carbon transferred from the atmosphere to soils (Fig. 3). The carbon-incorporation profiles can be compared with existing profiles of root biomass and above-ground inputs. The proportion of carbon that we found to be allocated to the subsoil is higher than the corresponding proportion of root biomass compiled in meta-analyses[5,26]. This can be explained on the one hand by the contribution of root exudates in addition to root mortality[27], and on the other hand by reduced decay rates at depth[24]. The reduced decay rates could be related to several interacting processes, for example, reduced and scattered microbial biomass[8], stabilization by minerals[8,18], and a reduced priming effect (the latter being the stimulation of SOC decomposition by the supply of fresh carbon)[28].

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

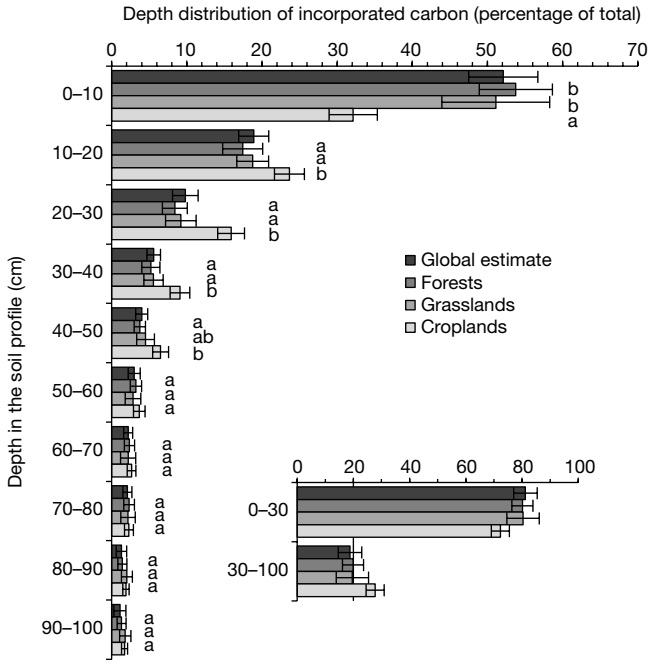


Fig. 3 | Depth distribution of the carbon that has been transferred from the atmosphere to soil organic matter between 1965 and 2015. The amount of carbon per 10-cm increment is expressed as a proportion of the total carbon incorporated in the top metre. The value for each land use is the mean of the observed profiles, and the value for the whole Earth was estimated by multivariate linear model extrapolation to the world's biomes. Error bars represent the 95% confidence interval of the mean or estimate; within each increment, land uses followed by the same letter (a or b) do not differ significantly. The small shift between the global estimate and the observed values reflects the differences in soil–climate conditions between the global average and the observation panel.

We measured the depth distribution of atmosphere-derived carbon incorporation over the past 50 years (50-year input). The depth distribution of the net change in soil carbon in the same time interval also depends on the loss of carbon older than 50 years during the period (50-year output). In steady-state systems, the depth distribution of outputs would perfectly equal the depth distribution of inputs. But real systems are transient as a result of global changes in either carbon inputs (for example, increased net primary production, reduced carbon returns because of land-use change) or decay rates (for example, because of warming). On the basis of our meta-analysis, we argue that the depth distributions of carbon output and of carbon incorporation are very similar even in transient systems, for the following reason. In non-steady-state systems, the delay associated with the downward movement of carbon may be suspected to result in 50-year outputs that are deeper than 50-year inputs, in a 'conveyor-like' dynamic system. But the $R_{30\text{-}100}$ ratio increases very slowly with time (by less than 0.001 per year; Extended Data Table 4). This means that the movement of carbon is slow and affects only long-term carbon dynamics, far later than the change expected in future decades. The depth distribution of net changes could differ from our distribution of new carbon only under the pressure of a driving force that affects old carbon in a very different way to the new carbon, such as de-freezing[29] or major changes in deep carbon inputs leading to additional priming effects[28].

Our study also reveals that the steep age gradient with depth (Fig. 2) could be a source of bias in the representation of carbon dynamics if depth is not handled properly. For instance, if we consider three commonly used reference layers—0–10 cm, 0–20 cm and 0–30 cm—we find that their median ages differ considerably, being 23, 50 and 92 years, respectively. Projecting the decay-rate parameters observed in the topmost part of soils onto thicker layers would bias future projections of changes in carbon. The kinetics of carbon incorporation further reveals a substantial turnover over the time range of centuries (Figs. 1, 2 and Extended Data Fig. 2)—that is, between the 'decadal' and 'millennial' compartments of present carbon models[1,7,30]—arguing for a more realistic description of carbon storage in terms of continuous time ranges[30].

Our results show that SOC dynamics and their responses to climatic control or land use are strongly depth dependent. A better representation of deep carbon dynamics has been called for, together with other processes, to improve ecosystem carbon models[7,12,19]. Our observations support the use of multilayer SOC modules in Earth system models, which our data could help to test.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0328-3

Received: 12 September 2017; Accepted: 4 May 2018;
Published online 11 July 2018.

1. Ahlström, A., Schurgers, G., Arneth, A. & Smith, B. Robustness and uncertainty in terrestrial ecosystem carbon response to CMIP5 climate change projections. *Environ. Res. Lett.* **7**, 044008 (2012).
2. Heimann, M. & Reichstein, M. Terrestrial ecosystem carbon dynamics and climate feedbacks. *Nature* **451**, 289–292 (2008).
3. Tiessen, H., Cuevas, E. & Chacon, P. The role of soil organic matter in sustaining soil fertility. *Nature* **371**, 783–785 (1994).
4. Rasmussen, P. E. et al. Long-term agroecosystem experiments: assessing agricultural sustainability and global change. *Science* **282**, 893–896 (1998).
5. Jobbágy, E. G. & Jackson, R. B. The vertical distribution of soil organic carbon and its relation to climate and vegetation. *Ecol. Appl.* **10**, 423–436 (2000).
6. Hiederer, R. & Köchy, M. *Global Soil Organic Carbon Estimates and the Harmonized World Soil Database* (Public. Office EU, 2011).
7. Todd-Brown, K. E. O. et al. Causes of variation in soil carbon simulations from CMIP5 Earth system models and comparison with observations. *Biogeosciences* **10**, 1717–1736 (2013).
8. Rumpel, C. & Kögel-Knabner, I. Deep soil organic matter—a key but poorly understood component of terrestrial C cycle. *Plant Soil* **338**, 143–158 (2011).
9. Carvalhais, N. et al. Global covariation of carbon turnover times with climate in terrestrial ecosystems. *Nature* **514**, 213–217 (2014).
10. Tian, H. Q. et al. Global patterns and controls of soil organic carbon dynamics as simulated by multiple terrestrial biosphere models: current status and future directions. *Glob. Biogeochem. Cycles* **29**, 775–792 (2015).
11. Le Quéré, C. et al. Global carbon budget 2016. *Earth Syst. Sci. Data* **8**, 605–649 (2016).
12. Luo, Y. et al. Toward more realistic projections of soil carbon dynamics by Earth system models. *Glob. Biogeochem. Cycles* **30**, 40–56 (2016).
13. Guan, X. K. et al. Soil carbon sequestration by three perennial legume pastures is greater in deeper soil layers than in the surface soil. *Biogeosciences* **13**, 527–534 (2016).
14. Hobley, E., Baldock, J., Hua, Q. & Wilson, B. Land-use contrasts reveal instability of subsoil organic carbon. *Glob. Change Biol.* **23**, 955–965 (2017).
15. Chen, G., Yang, Y. & Robinson, D. Allometric constraints on, and trade-offs in, belowground carbon allocation and their control of soil respiration across global forest ecosystems. *Glob. Change Biol.* **20**, 1674–1684 (2014).
16. Elzein, A. & Balesdent, J. Mechanistic simulation of vertical distribution of carbon concentrations and residence times in soils. *Soil Sci. Soc. Am. J.* **59**, 1328–1335 (1995).
17. Sierra, C. A., Müller, M., Metzler, H., Manzoni, S. & Trumbore, S. E. The muddle of ages, turnover, transit, and residence times in the carbon cycle. *Glob. Change Biol.* **23**, 1763–1773 (2017).
18. Mathieu, J., Hatté, C., Balesdent, J. & Parent, E. Deep soil carbon dynamics are driven more by soil type than by climate: a worldwide meta-analysis of radiocarbon profiles. *Glob. Change Biol.* **21**, 4278–4292 (2015).
19. He, Y. et al. Radiocarbon constraints imply reduced carbon uptake by soils during the 21st century. *Science* **353**, 1419–1424 (2016).
20. Ahrens, B. et al. Bayesian calibration of a soil organic carbon model using $\Delta^{14}C$ measurements of soil organic carbon and heterotrophic respiration as joint constraints. *Biogeosciences* **11**, 2147–2168 (2014).
21. Balesdent, J. & Mariotti, A. in *Mass Spectrometry of Soils* (eds Boutton, T. W. & Yamasaki, S. I.) 83–111 (Marcel Dekker, New York, 1996)
22. Lehmann, J. & Kleber, M. The contentious nature of soil organic matter. *Nature* **528**, 60–68 (2015).
23. Trabucco, A. & Zomer, R. *Global Aridity Index (Global-Aridity) and Global Potential Evapo-Transpiration (Global-PET) Geospatial Database* (CGIAR, Consortium for Spatial Information, 2009).
24. Guenet, B. et al. The relative importance of decomposition and transport mechanisms in accounting for soil organic carbon profiles. *Biogeosciences* **10**, 2379–2392 (2013).
25. Guo, L. & Gifford, R. Soil carbon stocks and land use change: a meta-analysis. *Glob. Change Biol.* **8**, 345–360 (2002).
26. Schenk, H. J. & Jackson, R. B. The global biogeography of roots. *Ecol. Monogr.* **72**, 311–328 (2002).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

27. Strand, A. E., Pritchard, S. G., McCormack, M. L., Davis, M. A. & Oren, R. Irreconcilable differences: fine-root life spans and soil carbon persistence. *Science* **319**, 456–458 (2008).
28. Fontaine, S. et al. Stability of organic carbon in deep soil layers controlled by fresh carbon supply. *Nature* **450**, 277–280 (2007).
29. Koven, C. D. et al. Permafrost carbon-climate feedbacks accelerate global warming. *Proc. Natl Acad. Sci. USA* **108**, 14769–14774 (2011).
30. Manzoni, S., Katul, G. G. & Porporato, A. Analysis of soil carbon transit times and age distributions using network theories. *J. Geophys. Res.* **114**, G04025 (2009).

Acknowledgements We thank C. Marol, S. Milin and P. Signoret for contributing to additional isotopic analyses, as well as the scientists who provided numerical data from their published studies. We thank the French Agence Nationale de la Recherche for funding through the projects Dedycas (14-CE01-0004) and Equipex Aster-CEREGE (ANR-10-EQPX-24) and for supporting the Institute National de la Recherche Agronomique (INRA) Laboratory UR-1138 through the Laboratory of Excellence ARBRE (ANR-11-LABX-0002-01). This is Laboratoire des Sciences du Climat et de l'Environnement contribution no. 6464.

Reviewer information *Nature* thanks I. Janssens and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.B. led the study and drafted the manuscript. All authors contributed equally to data provision and processing, and commented on and provided edits to the original manuscript. J.C. supervised the statistical analysis.

Competing interests The authors declare no financial competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0328-3.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0328-3.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.B.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

METHODS

Study sites. We compiled published datasets from 47 peer-reviewed articles, together with original data, on mineral soil $^{13}C/^{12}C$ changes in places where the $^{13}C/^{12}C$ ratio of the vegetation has been shifted for known durations (see Supplementary Information). We analysed a total of 112 pairs of mineral soil profiles: 108 pairs in which the predominant vegetation has changed from the C3 photosynthetic type to the C4 type, or vice versa, and four pairs from free-air carbon-enrichment (FACE) experiments, where the ^{13}C signature of added carbon dioxide has labelled plant-derived material (Extended Data Fig. 1; see references in the Supplementary Information). At each site, two plots with a common history (one with changed and one with unchanged vegetation) were analysed. The isotopic difference between the two profiles was used to calculate the proportion of new carbon through an isotope-mixing equation, which is not biased by additional isotopic effects in soils[21].

Most of the world's biomes are represented; the land uses include grasslands and savannahs (34%), forests and woodlands (30%), and annual and perennial crops (36%), from 24 countries between latitudes 29° S and 57° N (Extended Data Fig. 1). We selected studies that fulfil the following criteria: the age of change should be known or estimated; the observed depth should be more than 60 cm or reach bedrock; and the difference in the $\delta^{13}C$ of the vegetation between the reference and the study site should be 5‰ or more in the case of mixed vegetation that includes both photosynthetic types. The duration of vegetation change ranged from 4 years to 4,000 years. Authors estimated the dates of change through controlled experiments, enquiries, historical records, or airborne surveys. Changes in isotope signature that occurred more than 1,000 years ago (nine sites) were associated with interacting climate- and man-induced changes in vegetation. In those cases, dates were estimated by the authors from local or regional proxies of palaeovegetation change (for example, pollen/charcoal combined with radiocarbon dating). When the period after vegetation change was expressed by the authors as a range (for changes older than 200 years), we used the mid-value of the range.

Mean climatic data were obtained either from data reported in the article ($n = 103$) or, if missing ($n = 9$), from the CRU Group/Oxford/International Water Management Institute (IWMI) 10-minute mean climate grids for global land areas for the period 1961 to 1990 (ref. [31]). We compared grid versus declared climatic data in the database: for annual precipitation, the mean CRU grid/declared ratio is 0.98 ± 0.15 (standard deviation); for MAT, the mean difference between CRU grid and declaration is -0.15 ± 1.1 °C. Topsoil clay content was either obtained from authors' statements or assumed to be the median value of the texture class mentioned. Mean annual aridity indexes, P/PET (annual precipitation/potential evapotranspiration)[23]—a better indicator of hydric impact on both net primary production and microbial activity than precipitation alone—were obtained from the Food and Agriculture Organization 10-minute mean climate grids for global land areas for the period 1950 to 2000 (ref. [23]).

Proportion of new carbon and data pre-processing. For each site, the natural ^{13}C-labelling technique uses two plots, which were initially identical and have become differentiated during the last t years by two types of vegetation that differ in their $\delta^{13}C$. We use the terms 'reference' ('ref') for the plot at which the vegetation type at the date of sampling is the closest to that of the initial vegetation, and 'studied plot' ('s') for the plot with the new type of vegetation. Most authors described carbon content and isotopic data profiles as successive layers, each one sampled between two depths (z_1, z_2). For each layer (z_1, z_2), we define C as the carbon stock in the horizon (in kg m^{-2}); f_{new} as the proportion of new carbon (that is, derived from the new vegetation) (Fig. 1); and C_{new} as the stock of new carbon in the horizon (in kg m^{-2}). C, f_{new} and C_{new} were either obtained from the authors' papers ($n = 30$), or calculated from observed variables as follows. C was calculated from carbon concentration, [C] (in mg g^{-1}), and bulk density, ρ, according to $C = [C] \times \rho \times (z_2 - z_1)$, where ρ was either from the authors' data or (in 41 cases) estimated from [C] according to Alexander's[32] equation. f and C_{new} were calculated according to the equations[21]:

$$f_{new} = (\delta soil_s - \delta soil_{ref}) / \Delta\delta veg \quad (1)$$

$$C_{new} = f_{new} \times C$$

where $\delta soil_s$ and $\delta soil_{ref}$ are the $\delta^{13}C$ values of SOC from the study and reference plots; and $\Delta\delta veg$ is the difference in vegetation $\delta^{13}C$ between the study and reference plots and was determined from plant or litter samples. The $\delta soil_{ref}$ in each horizon was obtained from the reference soil collected at the same depth as the soil of the study plot. In accordance with the limit of resolution of the method, 27 horizons in deep layers had negative f_{new} values; in this case, we considered C_{new} to be 0. The resulting overestimation of average new carbon was negligible. In cases in which the sampling depth differed at the reference and studied plots, we calculated $\delta soil_s$ by linear interpolation of the two nearest observed depths. Equation (1) typically accounts for the various ^{13}C enrichments that occur during organic carbon decay or historical changes[21], with the sole criterion that these enrichments are similar in the study and reference soils. Equation (1) neglects the dark fixation of carbon atoms[33] that would have the isotopic composition of atmospheric CO_2.

Depth distribution of new carbon. We calculated depth distributions for the subset of sites whose labelling duration was 200 years or less ($n = 99$; Fig. 3). The mean duration was 35 years. In order to compare similar depth intervals, we calculated the three variables f_{new}, cumulative carbon stock with depth $C(0, z)$ and $C_{new}(0, z)$ 10-cm increments by linear interpolation of the observed horizons. For each 10-cm depth interval (z, $z + 10$ cm), we computed the ratio $R = C_{new}(z, +10\text{ cm})/C_{new}(0, 100\text{ cm})$. When bedrock or the R horizon was described, a nil carbon content was attributed to these horizons. When profiles were not described down to a depth of one metre ($n = 31$; most often 80 cm), $C(0, 100\text{ cm})$ was extrapolated from the maximum depth z_{max} using the linear regression $C(0, 100\text{ cm}) = a \times C(0, z_{max}) + b$ over the entire dataset and similarly for $C_{new}(0, 100\text{ cm})$.

The median depth z_{median} of new carbon was calculated for individual profiles as $C_{new}(0, z_{median}) = C_{new}(0, 100\text{ cm})/2$, by linear interpolation in the observed $C_{new}(0, z)$ function.

The variance of the ten ratios $R = C_{new}(z, z + 10\text{ cm})/C_{new}(0, 100\text{ cm})$ at the ten depths $z = 0, 10, \ldots, 90$ cm, the ratio for the whole subsoil $C_{new}(30, 100\text{ cm})/C_{new}(0, 100\text{ cm})$ and the median depth of new carbon were analysed by multivariate linear regression of time, land use and climatic variables (Extended Data Tables 4 and 5). Given that the average start date of labelling was 1965, we consider that the regression value of R for time = 50 years stands for carbon incorporated in the time interval 1965–2015. World average values of carbon incorporation in deep soil layers, excluding permafrost areas, were obtained by weighting multivariate linear regression estimates of new carbon (Extended Data Tables 1, 2, 4 and 5) by the biome soil carbon inventories in ref. [5]. Multivariate linear regression used the mean value of each of the 112 observed profiles, with no weighting for the number of replicates or horizons, leading to less precise but unbiased estimation. When replicated, profile variability is provided in the database in the Supplementary Information. We used bootstrap procedures[34] to express confidence on the estimated depth distribution or median age for the globe (Fig. 3 and Extended Data Table 6), or on the depth distribution of ages in tropical grasslands and forests (Fig. 2 and Extended Data Table 3). For that purpose, we drew $N = 100{,}000$ independent profile bootstrap samples from the observed profiles. For each bootstrap sample, relationships with P/PET, MAT, land use and time were recomputed and used to calculate the values of the variables of interest. Standard deviations were then estimated as the standard deviation of these 100,000 values.

Statistical analyses were performed using the R packages Boot and Stats version 3.4.3.

Analysis of the inference of vegetation change on the results. The naturally labelled sites experienced varying degrees of perturbation compared with pristine ecosystems. Vegetation change may modify input or decay rates, leading to transient carbon dynamics. To investigate whether these changes themselves affect the depth distribution of new carbon, we tested the dependence on two additional variables that characterize the observed sites: the previous type of vegetation—either crops, grassland or forest, known for 109 sites—and the relative difference in carbon stock between study and reference plots, when known and when the reference resembled the previous vegetation type ($n = 88$ sites). The relative change ΔC_{rel} is calculated as:

$$\Delta C_{rel} = [C(0, 100\text{ cm})_{studied\ site} - C(0, 100\text{ cm})_{reference\ site}]/C(0, 100\text{ cm})_{reference\ site}$$

ΔC_{rel} is nil on average in the database, that is, it corresponds to the steady state ($\Delta C_{rel} = 0.004 \pm 0.026$, ± s.e.m.); however, it does vary as a result of changes in inputs or dynamics in different directions. Mean durations of change are independent of previous vegetation in the statistical analysis: 31 years for previous grassland, 37 years for previous crops and 40 years for previous forests (excluding durations of more than 1,000 years, which involve no crop). ΔC_{rel} is not correlated with the duration of the change either.

Concerning the depth distribution of new carbon, that is, $R_{30\text{-}100} = C_{new}(30\text{ to }100\text{ cm})/C_{new}(0\text{ to }100\text{ cm})$, $R_{30\text{-}100}$ is not correlated with ΔC_{rel} either in the whole dataset ($r^2 = 0.002$; $n = 88$), or within the subsets of crops ($r^2 = 0.01$; $n = 31$), grasslands ($r^2 = 0.02$; $n = 24$) or forests ($r^2 = 0.13$; $n = 33$). We also tested the previous vegetation type as an explanatory variable of $R_{30\text{-}100}$ in addition to the other variables of climate, present land use and time (that is, the variables in Extended Data Table 4). The additional variable was not a significant factor (previous forest versus previous crop: $P = 0.88$; previous grassland versus previous crop: $P = 0.52$; previous grass versus previous forest: $P = 0.47$) and did not improve the model.

Concerning the proportion of new carbon in either topsoil or subsoil (that is, f_{new}), the previous vegetation type added as an explanatory variable in the statistical models of Extended Data Tables 1 and 2 was not a significant factors either ($P = 0.49$ to 0.99). By contrast, as an additional variable, ΔC_{rel} was highly significant for topsoil ($P < 0.01$) but was not for subsoil ($P = 0.12$). The effect is obvious given that both carbon change and new carbon are first driven by the relative change in inputs. This effect typically explains one of the results, namely the lower proportion of new carbon in cropland topsoils (Extended Data Table 1).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Concerning the age distribution in the subset of tropical grasslands and forests (Fig. 2 and Extended Data Table 3), the mean value of ΔC_{rel} is very low (0.02 ± 0.03; $\pm$ s.e.m.), close to steady state, and ΔC_{rel} does not depend on time, and therefore does not affect the estimated mean age distribution, but does contribute to the random dispersion of results.

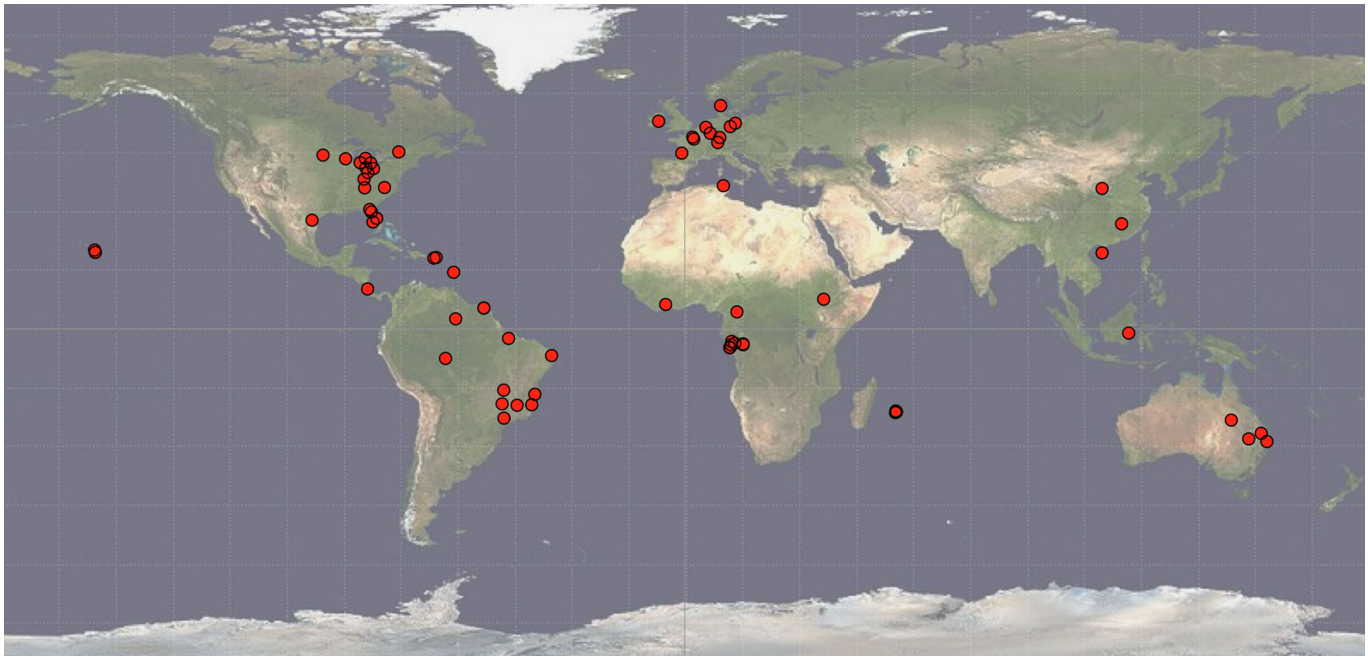
Finally, we included neither previous vegetation as an explanatory variable in the statistical models of the proportion of new carbon, nor carbon change, because of the covariance of ΔC_{rel} with land use. Furthermore, sites with previous or present croplands may have experienced a complex land-use history involving ancient primary forests and possibly pasture events. Taking all land-use histories into account would become a case-by-case study.

On the basis of this analysis of the inference of vegetation changes, we conclude that perturbation did not bias our estimates of the mean depth distribution of new carbon; that is, this depth distribution depends on the present vegetation and conditions, and not on previous vegetation, nor is it affected by non-steady-state conditions, in any systematic direction. The impact of perturbation on the proportion of new carbon in topsoils nevertheless prevented us from integrating our data towards global estimates of the absolute amount of new carbon or global carbon turnover. We thus restricted global integration to the depth distribution and median depth of new carbon.

Data availability. The raw primary data, calculated data and ancillary information analysed and generated here are available in the INRA public repository (https://doi.org/10.15454/KMNR6R). No statistical methods were used to predetermine sample size.

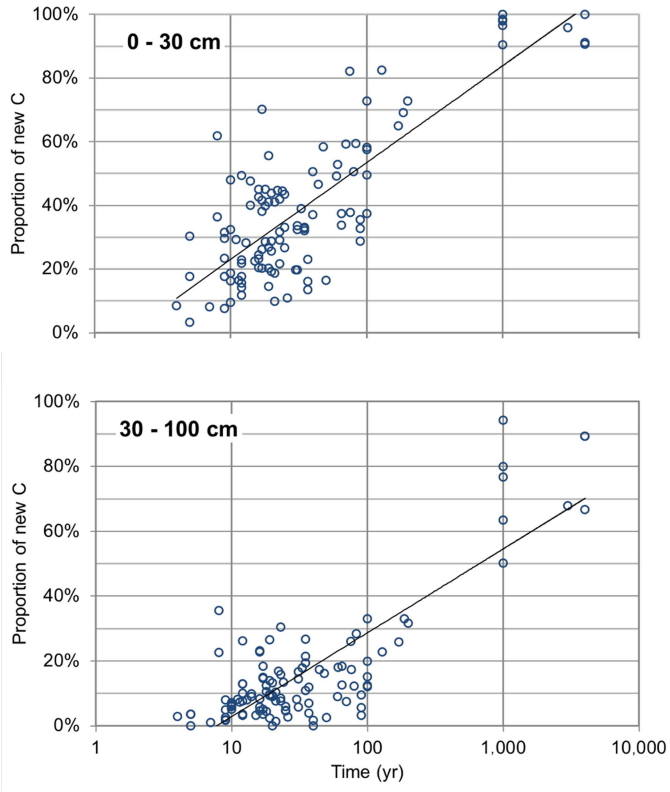
31. New, M., Lister, D., Hulme, M. & Makin, I. A high-resolution data set of surface climate over global land areas. *Clim. Res.* **21**, 1–25 (2002).
32. Alexander, E. B. Bulk densities of California soils in relation to other soil properties. *Soil Sci. Soc. Am. J.* **44**, 689–692 (1980).
33. Šantrůčková, H. et al. Significance of dark CO_2 fixation in arctic soils. *Soil Biol. Biochem.* **119**, 11–21 (2018).
34. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (Chapman and Hall, Boca Raton, 1993).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 1 | Locations of the study sites. Source of background image: Visible Earth, NASA.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 2 | Kinetics of new-carbon incorporation for the depth layers 0–30 cm and 30–100 cm. The respective logarithmic regressions $y = 0.30 \times \log_{10}(x) - 0.07$ for 0–30 cm and $y = 0.26 \times \log_{10}(x) - 0.23$ for 30−100 cm indicate that the duration required to replace one-third of the carbon is on average seven times longer in the subsoil than the topsoil.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 1 | Proportion of new carbon in topsoil: multivariate linear regression

	Coefficient estimate	Standard error	*T* value	Pr(>|*T*|)	
Intercept = Forest	-0.00250	0.0728	-0.034	0.973	
Grassland	-0.0531	0.0405	-1.311	0.193	
Cropland	-0.0951	0.0328	-2.896	0.00472	**
Log10(time)	0.258	0.038	6.865	7.6e-10	***
MAT	0.00595	0.00234	2.540	0.0128	*
P/PET	-0.0441	0.0337	-1.306	0.1954	
Clay	-0.000566	0.000693	-0.816	0.417	

The dependent variable is the ratio of new carbon (derived from the vegetation after time t) to total organic carbon in the topsoil layer. The explanatory variables are land use (grassland or cropland), $\log_{10}(t)$ (in years), mean annual temperature (MAT, in °C), ratio of annual precipitation to evapotranspiration (P/PET), and topsoil clay content (as a percentage). The reference land use (intercept) is forest. T is the value of Student's statistics; Pr(>|T|) is the probability value of the Student's t-test.
*$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.
Residual standard error, 0.1249 on 92 degrees of freedom; multiple R^2, 0.4781; adjusted R^2, 0.444; F-statistic, 14.04 on 6 and 92 degrees of freedom; P-value, 2.768×10^{-11}.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 2 | Proportion of new carbon in subsoil: multivariate linear regression

	Coefficient estimate	Standard error	*T* value	Pr(>|*T*|)	
Intercept = Forest	0.0205	0.0433	0.474	0.637	
Grassland	-0.0136	0.0241	-0.564	0.574	
Cropland	0.0024	0.0195	0.121	0.904	
Log10(time)	0.0849	0.0223	3.806	0.00025	***
MAT	0.00216	0.00139	1.551	0.124	
P/PET	-0.0516	0.0201	-2.570	0.0118	*
Clay	0.000018	0.000412	0.045	0.965	

The dependent variable is the ratio of new carbon (derived from the vegetation after time *t*) to total organic carbon in the subsoil layer. See Extended Data Table 1 for further definitions. Residual standard error, 0.07424 on 92 degrees of freedom; multiple R^2, 0.2561; adjusted R^2, 0.2076; *F*-statistic, 5.279 on 6 and 92 degrees of freedom; *P*-value, 0.0001028.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 3 | Age distribution of carbon over 55 tropical grassland and forest soil profiles

Depth	a_1	k_1 (yr^{-1})	a_2	k_2 (yr^{-1})	$a_1 + a_2$	Standard deviation of residuals
0 cm	0.614	0.21	0.34	0.0073	0.95	0.18
	(0.47, 0.75)	*(0.11, 2.8)*	*(0.2, 0.49)*	*(0.0024, 0.0158)*	*(0.92, 1.0)*	
10 cm	0.287	0.15	0.67	0.0059	0.96	0.11
	(0.19, 0.40)	*(0.08, 0.5)*	*(0.56, 0.77)*	*(0.0028, 0.0083)*	*(0.94, 1.0)*	
20 cm	0.108	0.23	0.85	0.0047	0.95	0.09
	(0.05, 0.19)	*(0.09, 1.5)*	*(0.78, 0.92)*	*(0.0025, 0.0072)*	*(0.92, 1.0)*	
30 cm	0.074	0.24	0.86	0.0035	0.93	0.10
	(0.02, 0.13)	*(0.13, 1.5)*	*(0.8, 0.93)*	*(0.0020, 0.0064)*	*(0.88, 1.0)*	
40 cm	0.070	0.23	0.83	0.0026	0.90	0.10
	(0.03, 0.13)	*(0.09, 1.0)*	*(0.74, 0.93)*	*(0.0014, 0.0041)*	*(0.83, 1.0)*	
50 cm	0.066	0.21	0.80	0.0020	0.87	0.09
	(0.04, 0.11)	*(0.10, 0.5)*	*(0.70, 0.95)*	*(0.0010, 0.0030)*	*(0.79, 1.0)*	
60 cm	0.065	0.20	0.75	0.0018	0.82	0.10
	(0.03, 0.11)	*(0.10, 0.4)*	*(0.65, 0.89)*	*(0.0008, 0.0028)*	*(0.73, 0.94)*	
70 cm	0.052	0.22	0.71	0.0016	0.76	0.10
	(0.02, 0.09)	*(0.13, 0.5)*	*(0.62, 0.88)*	*(0.0007, 0.0024)*	*(0.67, 0.95)*	
80 cm	0.044	0.25	0.65	0.0016	0.70	0.11
	(0.01, 0.08)	*(0.15, 3.2)*	*(0.54, 0.92)*	*(0.0005, 0.0024)*	*(0.60, 0.93)*	
90 cm	0.042	0.25	0.60	0.0016	0.64	0.11
	(0.01, 0.08)	*(0.14, 3.5)*	*(0.46, 0.99)*	*(0.0003, 0.0025)*	*(0.51, 1.0)*	
100 cm	0.048	0.25	0.55	0.0014	0.60	0.11
	(0.01, 0.08)	*(0.14, 3.3)*	*(0.44, 0.99)*	*(0.0002, 0.0025)*	*(0.48, 1.0)*	

These data were used to generate Fig. 2. At each depth, the proportion f_{new} of carbon aged less than t was fitted by a nonlinear regression of time t using the equation $f_{new} = a_1.[1 - \exp(-k_1 \times t)] + a_2.[1 - \exp(-k_2 \times t)]$. Such bi-exponential functions[30] describe carbon age distribution, with carbon divided into three age classes, a_1 being the proportion of 'young' carbon, a_2 the proportion of 'old' carbon, and $(1 - a_1 - a_2)$ the proportion of carbon with an infinite age. $1/k_1$ and $1/k_2$ are the mean ages of young and old carbon, respectively. Numbers in parentheses denote the 95% confidence intervals of the estimated parameters. The median environmental conditions of the soil set are: MAT = 23.6 °C; annual precipitation = 2,100 mm; P/PET (ref. [23]) = 1.44 and topsoil clay content = 37%.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 4 | Depth incorporation of new carbon in subsoil: multivariate linear regression

	Coefficient estimate	Standard error	*T* value	Pr(>|*T*|)	
Intercept = Forest	0.193	0.049	3.979	0.00014	***
Grassland	0.0218	0.03499	0.624	0.534	
Cropland	0.105	0.029	3.646	0.00044	***
Time (yr)	0.000369	0.000320	1.155	0.251	
MAT (°C)	0.00381	0.00206	1.848	0.0677	
P/PET	-0.0996	0.0297	-3.354	0.00116	**
Clay (%)	0.000638	0.000608	1.049	0.297	

The dependent variable is the ratio of the amount of new carbon (derived from the vegetation after time t, in kg m^{-2}) in the subsoil layer to the amount of new carbon in the entire top metre—that is, $R_{30\text{-}100} = C_{new}$(30 to 100 cm)/$C_{new}$(0 to 100 cm). See Extended Data Table 1 for further definitions. Note the dependence on time: the maximum value of the coefficient at the 95% confidence level (estimate + 2 s.e.m.) is 0.001 yr^{-1}. Residual standard error, 0.11 on 92 degrees of freedom; multiple R^2, 0.2387; adjusted R^2, 0.1891; F-statistic, 4.808 on 6 and 92 degrees of freedom; P-value, 0.0002618.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 5 | Median depth of new carbon: multiple linear regression

	Coefficient estimate	Standard error	*T* value	Pr(>|*T*|)	
Intercept = Forest	11.3	2.183	5.174	1.31e-06	***
Grassland	3.93	1.463	2.689	0.00849	**
Cropland	8.77	1.28	6.842	8.18e-10	***
Time (yr)	0.0201	0.0137	1.466	0.146	
MAT (°C)	0.183	0.089	2.045	0.0437	*
P/PET	-5.11	1.32	-3.871	0.000202	***

The dependent variable is the median depth (in cm) of the amount of new carbon (carbon derived from the vegetation after time t, in kg m^{-2}) of each profile. See Extended Data Table 1 for further definitions. Residual standard error, 4.963 on 93 degrees of freedom; multiple R^2, 0.3985; adjusted R^2, 0.3662; F-statistic, 12.32 on 5 and 93 degrees of freedom; P-value: 3.551×10^{-9}.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Table 6 | Depth distribution of carbon transferred from atmosphere to SOM in 1965–2015

	Observed Forests	Observed Grasslands	Observed Croplands	Global estimate
0-10 cm	0.538	0.511	0.321	0.521
	(0.048)	*(0.071)*	*(0.032)*	*(0.046)*
10-20 cm	0.174	0.188	0.237	0.189
	(0.026)	*(0.021)*	*(0.019)*	*(0.020)*
20-30 cm	0.084	0.092	0.159	0.098
	(0.016)	*(0.020)*	*(0.017)*	*(0.017)*
30-40 cm	0.052	0.056	0.091	0.056
	(0.012)	*(0.012)*	*(0.013)*	*(0.009)*
40-50 cm	0.037	0.045	0.065	0.040
	(0.007)	*(0.011)*	*(0.010)*	*0.008)*
50-60 cm	0.032	0.028	0.037	0.030
	(0.007)	*(0.010)*	*(0.007)*	*0.008)*
60-70 cm	0.024	0.022	0.026	0.022
	(0.006)	*(0.010)*	*(0.005)*	*0.006)*
70-80 cm	0.023	0.022	0.023	0.021
	(0.007)	*(0.009)*	*(0.005)*	*(0.006)*
80-90 cm	0.014	0.020	0.019	0.013
	(0.006)	*(0.007)*	*(0.004)*	*(0.007)*
90-100 cm	0.013	0.018	0.017	0.011
	(0.005)	*(0.007)*	*(0.004)*	*(0.008)*

These data were used to generate Fig. 3. The amount of new carbon transferred from the atmosphere to SOM in 1965 to 2015 (<50 yr carbon) in each 10-cm layer is expressed as a proportion of the total new carbon <50 yr in the first metre. The data shown are mean values for observed forests, grasslands and croplands, and global estimates; the numbers in parentheses are the 95% confidence intervals on the mean or estimate.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0335-4

Rapid glaciation and a two-step sea level plunge into the Last Glacial Maximum

Yusuke Yokoyama[1,2,3]*, Tezer M. Esat[4,5], William G. Thompson[6], Alexander L. Thomas[7], Jody M. Webster[8], Yosuke Miyairi[1], Chikako Sawada[1], Takahiro Aze[1], Hiroyuki Matsuzaki[9], Jun'ichi Okuno[10], Stewart Fallon[4], Juan-Carlos Braga[11], Marc Humblet[12], Yasufumi Iryu[13], Donald C. Potts[14], Kazuhiko Fujita[15], Atsushi Suzuki[16] & Hironobu Kan[17]

The approximately 10,000-year-long Last Glacial Maximum, before the termination of the last ice age, was the coldest period in Earth's recent climate history[1]. Relative to the Holocene epoch, atmospheric carbon dioxide was about 100 parts per million lower and tropical sea surface temperatures were about 3 to 5 degrees Celsius lower[2,3]. The Last Glacial Maximum began when global mean sea level (GMSL) abruptly dropped by about 40 metres around 31,000 years ago[4] and was followed by about 10,000 years of rapid deglaciation into the Holocene[1]. The masses of the melting polar ice sheets and the change in ocean volume, and hence in GMSL, are primary constraints for climate models constructed to describe the transition between the Last Glacial Maximum and the Holocene, and future changes; but the rate, timing and magnitude of this transition remain uncertain. Here we show that sea level at the shelf edge of the Great Barrier Reef dropped by around 20 metres between 21,900 and 20,500 years ago, to −118 metres relative to the modern level. Our findings are based on recovered and radiometrically dated fossil corals and coralline algae assemblages, and represent relative sea level at the Great Barrier Reef, rather than GMSL. Subsequently, relative sea level rose at a rate of about 3.5 millimetres per year for around 4,000 years. The rise is consistent with the warming previously observed at 19,000 years ago[1,5], but we now show that it occurred just after the 20-metre drop in relative sea level and the related increase in global ice volumes. The detailed structure of our record is robust because the Great Barrier Reef is remote from former ice sheets and tectonic activity. Relative sea level can be influenced by Earth's response to regional changes in ice and water loadings and may differ greatly from GMSL. Consequently, we used glacio-isostatic models to derive GMSL, and find that the Last Glacial Maximum culminated 20,500 years ago in a GMSL low of about −125 to −130 metres.

The sea level rise between the Last Glacial Maximum (LGM) and the Holocene was at times episodic and stimulated particular patterns in coral reef growth and evolution. Estimates of the maximum volume of excess ice and amount of water that contributed to the change in the corresponding GMSL was originally based on the stratigraphy of radiocarbon and U-series dated corals from Barbados[6,7]. Data from radiocarbon-dated microfossils and macrofossils also helped to define LGM palaeo-shorelines from the Sunda Shelf in the South China Sea[8] and the Bonaparte Gulf of northern Australia[5,9]. However, large uncertainties remain in local relative sea level (RSL) data and in glacio-isostatic (GIA) model inputs, such as ice histories and Earth rheology. These are needed to estimate the bounds of past and future GMSLs, which, for the LGM, range[4,10] from 115 m to 135 m.

Precise and accurate sea level histories, often derived from dating of fossil corals and algae, are important. Coastal shelf geometry and location of land-bridges and islands at the LGM have a bearing on probable human migration routes and affect the ecological and species diversity, for example, of endemic flowering plants on islands that had complex, dynamic histories and covered a larger area during the LGM[11]. Sea levels vary with polar ice volumes and the extent of LGM ice-sheets can affect atmospheric pressure patterns and alter the salinity of oceans causing circulation changes[4].

Fossil coral and coralline algae deposits of the LGM–Holocene period at the Great Barrier Reef (GBR) are below the present sea level and were drilled during the Integrated Ocean Drilling Program (IODP) Expedition 325 in 2010[12] (Fig. 1). Corals and coralline algae were recovered from 34 holes, over two transects 500 km apart, at Hydrographer's Passage (HYD-01C) and Noggin Pass (NOG-01B; Fig. 1). Depths of up to 150 m below sea level were reached to access the full LGM period[12]. Selected well characterized samples were dated using U-series (coral) and accelerator-based radiocarbon (coral and algae) methods (Extended Data Table 1). Sea level depth uncertainties depend on the palaeo-habitat depth range of particular coral and algae species and were conservatively assessed in conjunction with associated algal crust thickness, vermetid gastropods and by benthic foraminiferal assemblages[12,13] (Figs. 2 and 3, Methods and Extended Data Figs. 1–6).

A brief outline of previous determinations of the timing and duration of the LGM shows that it extended from about 29,500 years (29.5 kyr) ago to 19 kyr ago[1]. There is an initial rapid (>40 m) fall in GMSL from 31–32 kyr ago to 29–30 kyr ago (Fig. 4a, b)[1,4]. A protracted gradual GMSL drop was construed at about 29–21 kyr ago[4]. However, this was largely an extrapolation between the two endpoints owing to the sparsity of data which also have large (about 20 m) uncertainties in RSL elevations (Fig. 4c, e)[4–7]. The onset of deglaciation is apparent from 21 kyr ago with a gradual 10–15 m GMSL rise[4] followed by a short stable, or possibly slowly falling, GMSL from around 18 kyr ago to 16.5 kyr ago (see figure 4a, b of ref. [4]). From then on, the deglaciation proceeded at a fast pace, at times exceeding 12 m kyr^{-1} during the so-called meltwater pulses[14,15].

We converted our new GBR local sea levels to global values through GIA modelling (Methods), which accounts for the higher GBR coastline elevations owing to increased ice volumes and reduced adjacent ocean water loading. Our new GBR local sea levels, converted to global values through GIA modelling (Methods), substantially diverge from earlier determinations and completely revise the internal structure of the LGM GMSL (Fig. 4b). A large number of data points from the Noggin Pass and Hydrographer's Passage sections show a relatively constant sea level from about 28 kyr ago to 22 kyr ago after the initial rapid fall from 31 kyr ago[4], as documented previously from Barbados, Huon and the Bonaparte Gulf (Figs. 3 and 4c–h). The GMSL remains invariant from

[1]Atmosphere and Ocean Research Institute, University of Tokyo, Kashiwa, Japan. [2]Department of Earth and Planetary Science, Graduate School of Science, University of Tokyo, Tokyo, Japan. [3]Japan Agency for Marine-Earth Science and Technology, Yokosuka, Japan. [4]Research School of Earth Sciences, Australian National University, Canberra, Australian Capital Territory, Australia. [5]Research School of Physics and Engineering, Australian National University, Canberra, Australian Capital Territory, Australia. [6]Woods Hole Oceanographic Institution, Woods Hole, MA, USA. [7]University of Edinburgh, Edinburgh, UK. [8]University of Sydney, Sydney, New South Wales, Australia. [9]University Museum, University of Tokyo, Tokyo, Japan. [10]National Institute of Polar Research, Tokyo, Japan. [11]Universidad de Granada, Granada, Spain. [12]Nagoya University, Nagoya, Japan. [13]Tohoku University, Sendai, Japan. [14]University of California Santa Cruz, Santa Cruz, CA, USA. [15]University of Ryukyu, Naha, Japan. [16]Geological Survey of Japan, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan. [17]Kyushu University, Fukuoka, Japan. *e-mail: yokoyama@aori.u-tokyo.ac.jp

© 2018 Springer Nature Limited. All rights reserved.

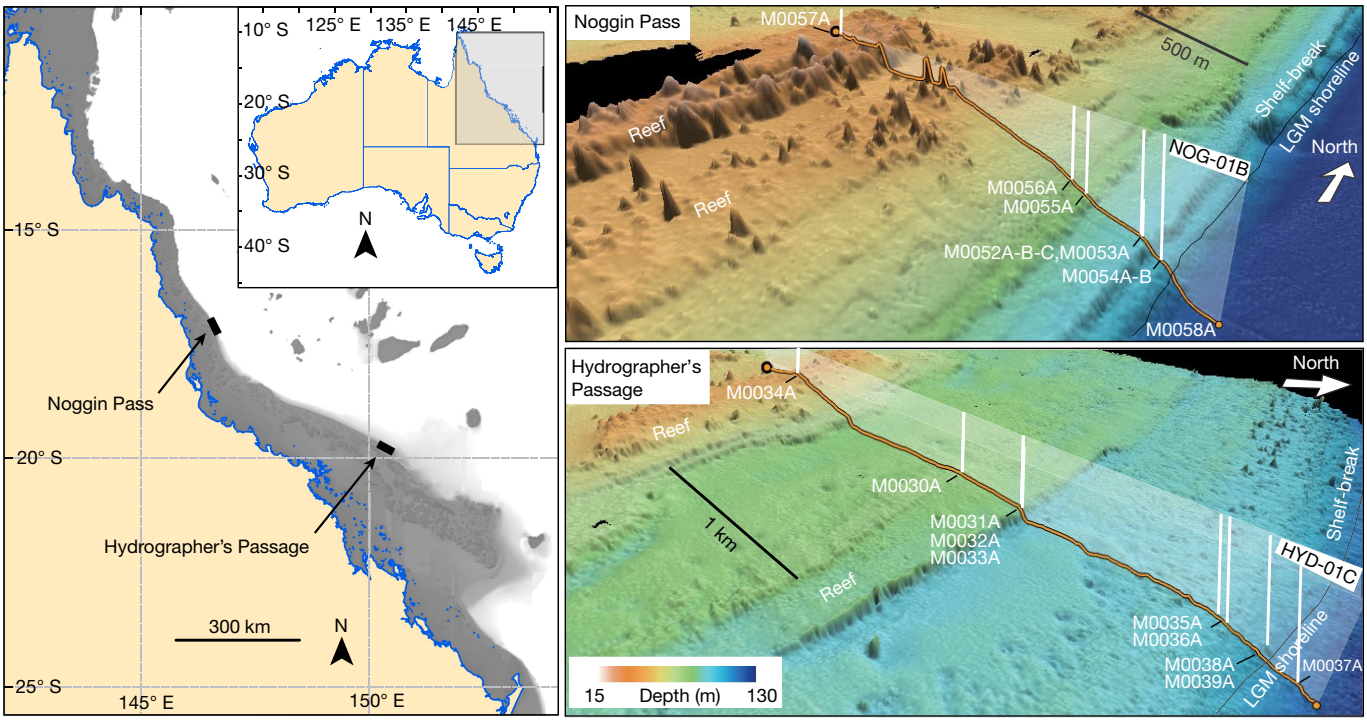


Fig. 1 | Location of GBR Expedition 325 study site. Maps show the northern and southern sites at Cairns (Noggin Pass, NOG-01B) and at Mackay (Hydrographer's Passage, HYD-01C). The high-resolution three-dimensional multibeam images show the surface geomorphic context[12], drill transects and specific locations of the drill holes.

30 kyr ago to 21.5 kyr ago at a mean value of 113 m within a range of about 6 m and represents the early LGM period, LGM-a (Fig. 4b; red band). This is much shallower than former estimates[4] (Fig. 4b). The previously identified[16,17] 'onset of deglaciation' about 19 kyr ago in fact corresponds to a rapid sea level fall followed by an approximately 4-kyr-long and 3.5 m kyr^{-1} sea level rise. We label this period LGM-b, lasting from about 21 kyr ago to 17 kyr ago. The minimum GMSL at LGM-b ranges from about 130 m to 125 m owing to analytical uncertainties and uncertainties in Earth model parameters and occurs after an approximately 20 m drop (Methods, Extended Data Figs. 8 and 9 and Extended Data Table 1). This point defines the start of LGM-b, after which the GMSL slowly rises to 120 m about 17 kyr ago followed by a rapid transition towards full deglaciation. The rates of net ice mass gain or sea level fall at the inception of both the LGM-a (about 30 kyr ago) and LGM-b (about 21 kyr ago) are similar, in the approximate range 15–20 mm yr^{-1}, as determined from the slopes of the GMSL curves. This GIA-modelled, very fast, maximum glaciation rate is notably faster than the mean LGM–Holocene deglaciation rate[4] of around

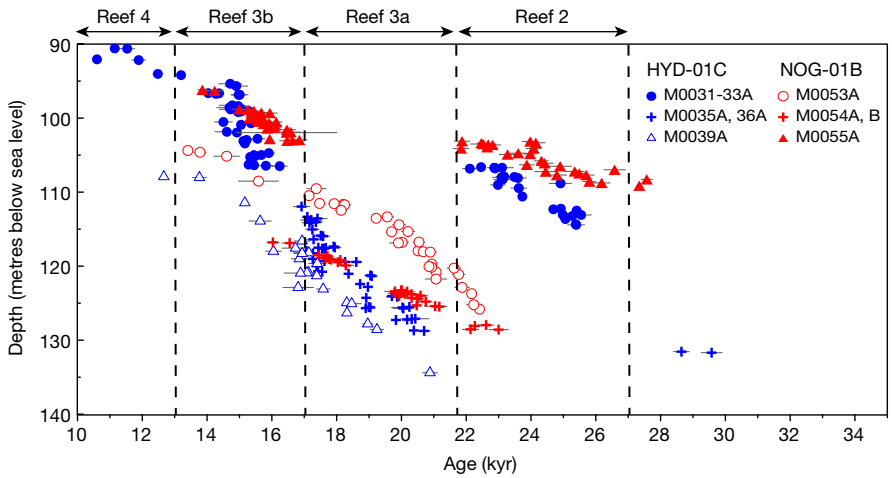


Fig. 2 | LGM sea level drop at LGM-b. Here, the sea level drop is apparent as a distinct age discontinuity observed in cores M0055A and M0031-33A from the NOG and HYD transects, respectively, separated by more than 500 km in the GBR. Cores obtained from different fossil GBR terraces and reefs reveal the trajectory of past sea level changes. Five reef sequences are distinguished on the basis of the IODP Expedition 325 record: Reef 1 (≥30 kyr old), Reef 2 (27–22 kyr old), Reef 3 (3a, 21–17 kyr old; 3b, 17–13 kyr old), Reef 4 (13–10 kyr old) and Reef 5 (the modern GBR). The major growth hiatus evident from both the HYD and the NOG sections marks the death of Reef 2 (22.1–21.9 kyr ago) following the sea level fall leading to LGM-b and re-establishment of reef (Reef 3a) further seaward 20.7–20.5 kyr ago (Extended Data Figs. 3–5). The age versus depth relationships, and the presence of freshwater low-magnesium calcite cement, in Reef 2, confirms subaerial exposure during the LGM-b period. In contrast, their absence below Reef 3a deposits places a maximum limit on the drop in sea level (Methods). At the end of the LGM, 17 kyr ago, the sea level rose rapidly, flooding the outer shelf and causing the re-establishment of the reef over its former position (Reef 3b), marking the end of the hiatus at the top of Reef 2 about 17 kyr ago. The details of the transition to LGM-b and the critical samples defining the fall in sea level are shown in Extended Data Fig. 6.

© 2018 Springer Nature Limited. All rights reserved.

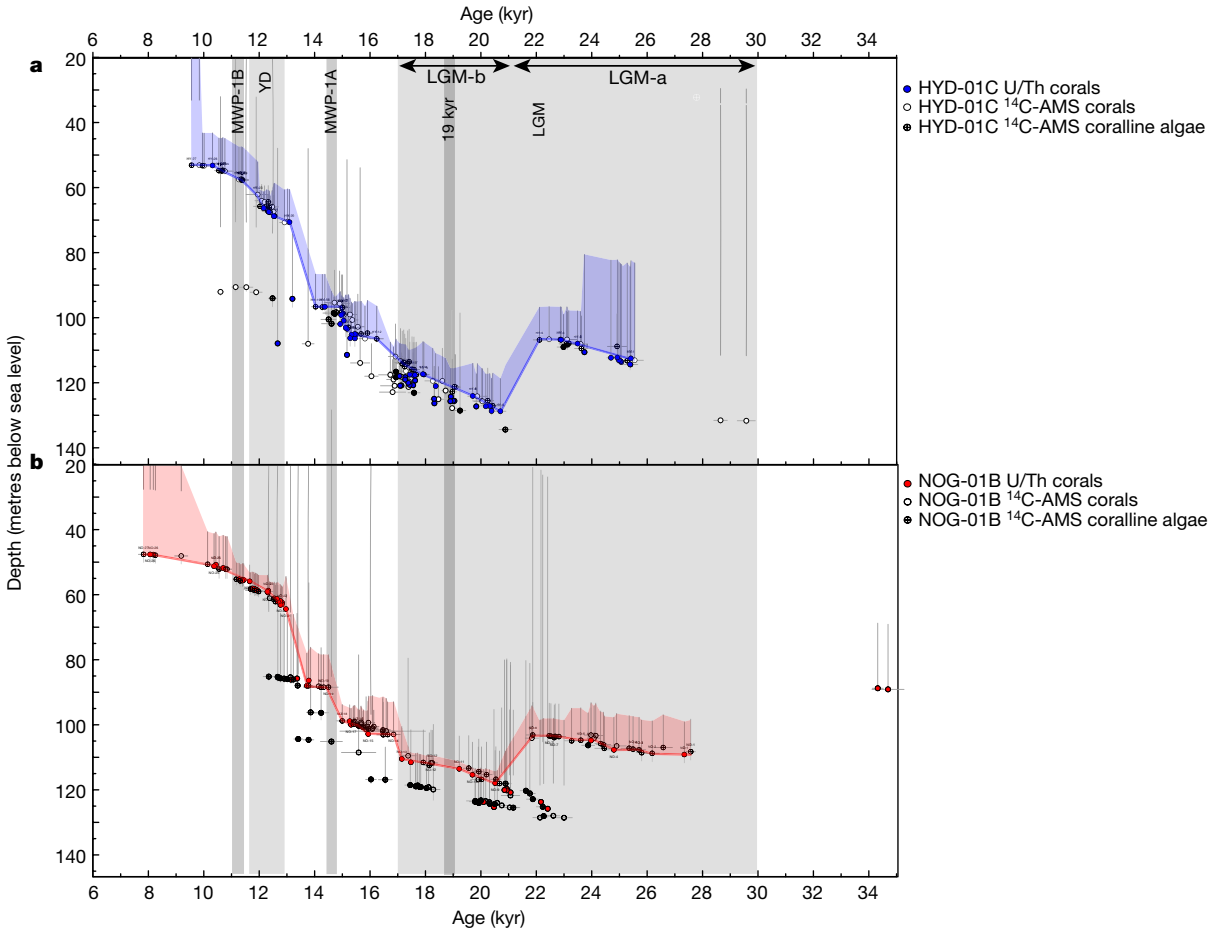


Fig. 3 | Age versus depth plots showing the RSL envelopes derived from samples recovered by IODP Expedition 325. a, **b**, Both HYD (**a**, blue) and NOG (**b**, red) transect cores were examined for coral, coralgal algae and benthic foraminiferal assemblages (Extended Data Figs. 1 and 2) with the help of X-ray computed tomography scanning and X-ray diffraction (Supplementary Information), which revealed detailed features of RSL histories during the past 35 kyr. The >500 dates, selected through detailed sedimentologic and biologic analyses (Methods), provided a robust chronostratigraphic framework that defined five distinct reef sequences (Reefs 2 to 4 are labelled in Fig. 2), which grew episodically over the past 30 kyr. The RSLs constructed here depend on the age (horizontal grey lines) and the sea level depth uncertainties (bars are ranges) related to palaeo-habitat depth range (grey upward lines) of each dated coral or algal sample and the maximum coring depth uncertainty (grey downward lines). The palaeowater depth ranges for shallow water species (blue and red shading) were conservatively estimated using a multiproxy approach combining coral, coralline algae and other key indicators such as algal crust thickness, and the presence of vermetid gastropods (Methods, Extended Data Table 1). We note that the disconnected palaeowater depth lines on some samples are indicative of deeper habitat ranges, probably >20 m water depth. The distribution of marine and freshwater cements in the cores were used to support the new estimates of the timing and magnitude of the LGM-b sea levels (Extended Data Figs. 3–5), including hiatuses and regressions.

12 mm yr^{-1}. The sea level drop at LGM-b is strikingly evident in cores from both HYD and NOG sections as a major growth hiatus in cores M0031-33A and M0055A-33 (Fig. 2). Deposits of freshwater calcite cement, in corals below the hiatus, indicate subaerial exposure and low sea levels (Methods; Extended Data Fig. 5). In turn, the gap corresponding to the low in sea level is picked up at five other sites, M0035, M0036, M0039, M0053 and M0054, at lower elevation, to complete the sea level curve (Extended Data Fig. 3).

We have established an ice model, based on the new GBR RSL data, that provides good agreement with other far-field RSL records when combined with a range of Earth model parameters that span the possible range of viscosities for the upper and lower mantle as well as lithospheric thickness. Model predictions of RSLs over the LGM-a and LGM-b periods generally agree well with the trends derived from existing data (Fig. 4, Extended Data Fig. 7; Supplementary Information). The magnitude of the sea level drop at the start of the LGM[1,4] (about 29 kyr ago) is about 40 m, constrained by coral data from the Huon Peninsula[18] and Barbados[19], and by foraminifera oxygen isotope records from the Red Sea[20] within relatively large (about 10 m) uncertainties. There is also close agreement between model results and data over the deglaciation period from 17 kyr ago onwards except at Barbados (Fig. 4c). Here, the systematic overshoot of the data above the sea level curve is indicative of additional processes, possibly tectonic in nature[15]. This is consistent with independent evaluations including a study[10] that uses the rate of change of degree-two harmonics of Earth's geopotential due to GIA.

The very rapid build-up of global ice volume during the two periods of transition at LGM-a (about 29–31 kyr ago; ref. [4], Fig. 4d) and LGM-b (about 22–21 kyr ago; Fig. 4b) requires substantial moisture transport and snow precipitation over existing ice sheets. To accommodate LGM-b an additional equivalent ice volume corresponding to up to 17 m of sea level is required. The location of the extra ice cannot be determined with certainty. GIA modelling and Northern and Southern Hemisphere bipolar climate paced by complementary high-latitude insolation highs[1,21,22] at these times shows increased ice volume over the North American Ice Sheet at the LGM, whereas the Eurasian ice sheet appears to have grown at a slower pace and commenced melting after about 22 kyr ago (Extended Data Fig. 8). Colder Antarctic climate during the LGM[3,16] is likely to have hindered ice calving and lessened basal melting of ice shelves, resulting in increased ice volume. The sustained growth of the Antarctic Ice Sheet during the LGM-b period and beyond, including continuing ice accumulation up to around 14 kyr

© 2018 Springer Nature Limited. All rights reserved.

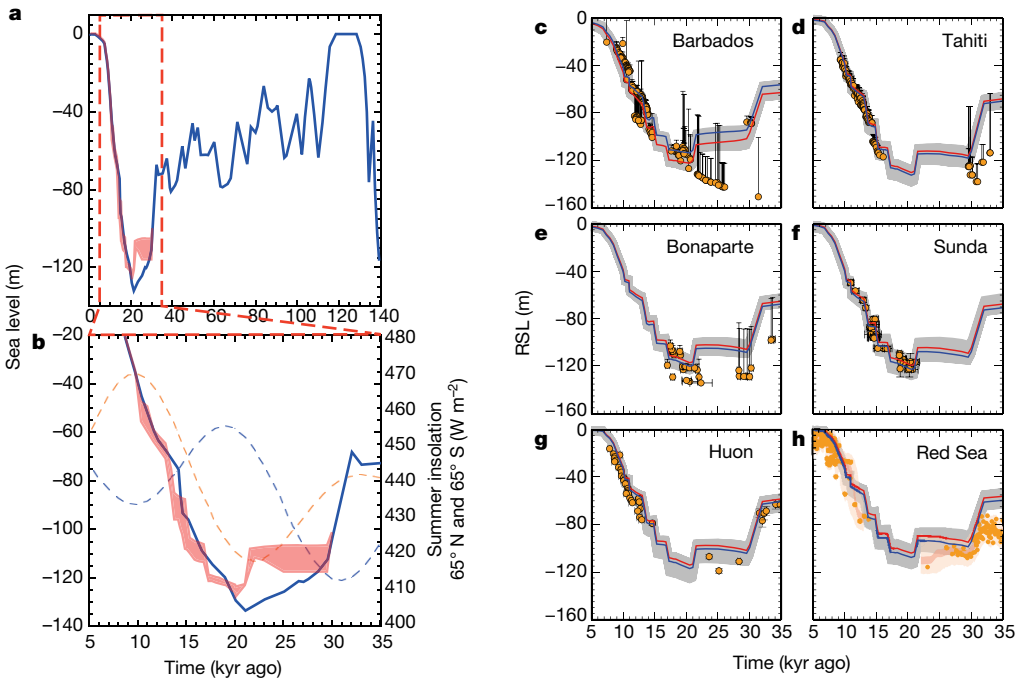


Fig. 4 | GMSL and GIA-model-derived far-field RSL with observational data. a–h, The previously constructed GMSL curve over the past 140 kyr (ref. [4]; blue line in **a** and **b**), is shown together with GIA model predictions, in **c**–**h**, for selected far-field data represented by red circles with depth and age error bars (standard deviation). High-latitude (65°) summer insolation curves for the Northern and Southern Hemispheres in **b** are shown as dashed orange and blue lines, respectively. Long-term sea level variation is in step with Northern Hemisphere summer insolation, whereas LGM-b occurs at the peak of Southern Hemisphere insolation. Orange circles in **c**–**h** indicate previously reported RSL with depth and age uncertainties (horizontal and vertical black lines). Results of 62 GIA model runs (Methods), over a range of potential Earth model parameters, such as lithospheric thickness and the viscosities of both the upper and lower mantle, are within the orange shading in **a** and **b** and grey shading in **c**–**h**. The red and blue curves in the grey shading in **c**–**h** show the results for lower-mantle viscosities of 10^{22} Pa s and 10^{23} Pa s, respectively, with a lithospheric thickness of 70 km and upper-mantle viscosity of 10^{20} Pa s. We note the foraminiferal oxygen-isotope-based Red Sea data[20] (**h**), which supports the new GMSL calculations from the present study within uncertainties (about 10 m). The orange shading in **h** represents confidence intervals of 95% for the RSL data (light orange) and the probability maximum (dark orange) reported in ref. [20].

ago, agrees with observations of the late retreat of the West Antarctic Ice Sheet at this time[1,23,24]. However, the major increase in ice volume, precipitating the onset of LGM-b, appears to have been during a short period (20 to 21 kyr ago) over the North American Ice Sheet, after which the North American Ice Sheet retreated from about 20 kyr ago onwards[25] (Extended Data Fig. 8).

The enlarged global ice volume about 30 kyr ago[1,4,18,26], equivalent to about 40 m of sea-level drop, persisted for over 8 kyr. Similarly, the low Northern Hemisphere insolation and somewhat reduced atmospheric carbon dioxide levels around 21–22 kyr ago led to a period of cold climate, very low tropical Atlantic sea surface temperatures and ultimately the transition to LGM-b. At this time, increased Southern Hemisphere insolation is likely to have facilitated moisture transport to the South, increasing the Antarctic Ice Sheet volume[1,18,27]. The Heinrich event 1 about 17 kyr ago marks the end of LGM-b when full deglaciation kicked in as the pace of Northern Hemisphere insolation and atmospheric carbon dioxide levels increased rapidly (Fig. 4).

The two sharp transitions preceding the LGM-a and LGM-b periods, associated with rapid accumulation of ice and lower sea levels, at the end of the last ice age, do not appear to be explicable in terms of processes attributable to any specific climate-change dynamic. During this time (around 29–19 kyr ago), oxygen isotope records in ice cores do not show a clear, distinct signal; the carbon dioxide levels were stable, insolation at the time was not very different to the present and tropical sea surface temperature did not change notably[1,28]. A systematic behaviour in sea level and climate has previously been noted[29] whereby transitions between two states, cold to warm or warm to cold, took place via a third state that was more extreme, as in the MIS3–LGM–Holocene transition (where MIS3 indicates the last ice age at marine isotope stage 3) or at the end of the last interglacial, a final high-stand before the MIS 5d transition. A similar behaviour in climate was previously noted where the appearance of an intermediate, extreme third state may have resulted in a shift from '41-kyr' cycles to '100-kyr' cycles about 800–1,000 kyr ago[30]. These bifurcations can be thought of as distinct energy states of climate between which 'stochastic perturbations' can cause transitions[30]. It appears that the same behaviour may occur over short timescales, not only over 100-kyr cycles. The transitions are not only manifest in climate but are also associated with sea level change.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0335-4.

Received: 6 April 2017; Accepted: 5 June 2018;
Published online 25 July 2018.

1. Clark, P. U. et al. The Last Glacial Maximum. *Science* **325**, 710–714 (2009).
2. Felis, T. et al. Intensification of the meridional temperature gradient in the Great Barrier Reef following the Last Glacial Maximum. *Nat. Commun.* **5**, 4102 (2014).
3. Mix, A. C., Bard, E. & Schneider, R. Environmental processes of the ice age: land, oceans, glaciers (EPILOG). *Quat. Sci. Rev.* **20**, 627–657 (2001).
4. Lambeck, K., Rouby, H., Purcell, A., Sun, Y. & Sambridge, M. Sea level and global ice volumes from the Last Glacial Maximum to the Holocene. *Proc. Natl Acad. Sci. USA* **111**, 15296–15303 (2014).
5. Yokoyama, Y., Lambeck, K., DeDeckker, P., Johnston, P. & Fifield, L. K. Timing of the Last Glacial Maximum from observed sea-level minima. *Nature* **406**, 713–716 (2000).
6. Fairbanks, R. G. A 17,000-year glacio-eustatic sea level record: influence of glacial melting dates on Younger Dryas event and deep ocean circulation. *Nature* **342**, 637–642 (1989).
7. Bard, E., Hamelin, B. & Fairbanks, R. G. U–Th ages obtained by mass spectrometry in corals from Barbados: sea level during the past 130,000 years. *Nature* **346**, 456–458 (1990).

© 2018 Springer Nature Limited. All rights reserved.

8. Hanebuth, T., Stattegger, K. & Grootes, P. M. Rapid flooding of the Sunda Shelf: a late-glacial sea-level record. *Science* **288**, 1033–1035 (2000).
9. De Deckker, P. & Yokoyama, Y. Micropalaeontological evidence for Late Quaternary sea-level changes in Bonaparte Gulf, Australia. *Global Planet. Change* **66**, 85–92 (2009).
10. Nakada, M., Okuno, J. & Yokoyama, Y. Total meltwater volume since the Last Glacial Maximum and viscosity structure of Earth's mantle inferred from relative sea level changes at Barbados and Bonaparte Gulf and GIA-induced J2. *Geophys. J. Int.* **204**, 1237–1253 (2016).
11. Weigelt, P., Steinbauer, M. J., Cabral, J. S. & Kreft, H. Late Quaternary climate change shapes island biodiversity. *Nature* **532**, 99–102 (2016).
12. Webster, J. M. et al. Response of the Great Barrier Reef to sea level and environmental changes over the past 30 ka. *Nat. Geosci.* **11**, 426–432 (2018).
13. Yokoyama, Y. & Esat, T. M. in *Handbook of Sea-Level Research* (eds Shennan, I., Long, A. & Horton, B.) Ch. 7, 104–124 (John Wiley & Sons, Oxford, 2015).
14. Deschamps, P. et al. Ice-sheet collapse and sea-level rise at the Bølling warming 14,600 years ago. *Nature* **483**, 559–564 (2012).
15. Bard, E., Hamelin, B. & Delanghe-Sabatier, D. Deglacial meltwater pulse 1B and Younger Dryas sea levels revisited with boreholes at Tahiti. *Nature* **327**, 1235–1237 (2010).
16. Clark, P. U., McCabe, A. M., Mix, A. C. & Weaver, A. J. Rapid rise of sea level 19,000 years ago and its global implications. *Science* **304**, 1141–1144 (2004).
17. MARGO Project Members. Constraints on the magnitude and patterns of ocean cooling at the Last Glacial Maximum. *Nat. Geosci.* **2**, 127–132 (2009).
18. Cutler, K. B. et al. Rapid sea-level fall and deep ocean temperature change since the last interglacial period. *Earth Planet. Sci. Lett.* **206**, 253–271 (2003).
19. Peltier, W. R. & Fairbanks, R. G. Global glacial ice volume and Last Glacial Maximum duration from an extended Barbados sea level record. *Quat. Sci. Rev.* **25**, 3322–3337 (2006).
20. Grant, K. M. et al. Rapid coupling between ice volume and polar temperature over the past 150,000 years. *Nature* **491**, 744–747 (2012).
21. Abe-Ouchi, A. et al. Insolation-driven 100,000-year glacial cycles and hysteresis of ice-sheet volume. *Nature* **500**, 190–193 (2013).
22. Philippon, G. et al. Evolution of Antarctic ice sheet throughout the last deglaciation: a study with a new coupled climate—north and south hemisphere ice sheet model. *Earth Planet. Sci. Lett.* **248**, 750–758 (2006).
23. Anderson, J. B. et al. Ross Sea paleo-ice sheet drainage and deglacial history during and since the LGM. *Quat. Sci. Rev.* **100**, 31–54 (2014).
24. Yokoyama, Y. et al. Widespread collapse of the Ross Ice Shelf during the late Holocene. *Proc. Natl Acad. Sci. USA* **113**, 2354–2359 (2016).
25. Lambeck, K., Purcell, A. & Zhao, S. The North American Late Wisconsin ice sheet and mantle viscosity from glacial rebound analyses. *Quat. Sci. Rev.* **158**, 172–210 (2017).
26. Clark, P. U. & Tarasov, L. Closing the sea level budget at the Last Glacial Maximum. *Proc. Natl Acad. Sci. USA* **111**, 15861–15862 (2014).
27. .Clark, P. U., Hostetler, S. W., Pisias, N. G., Schmittner, A. & Meissner, K. J. in *Ocean Circulation* (eds Schmittner, A., Chiang, J. C. H. & Hemming, S. R.) 209–246 (Geophysical Monograph 173, American Geophysical Union, Washington, 2007).
28. Brook, E. J. et al. Timing of millennial-scale climate change at Siple Dome, West Antarctica, during the last glacial period. *Quat. Sci. Rev.* **24**, 1333–1343 (2005).
29. Yokoyama, Y. & Esat, T. M. Global climate and sea level-enduring variability and rapid fluctuations over the past 150,000 years. *Oceanography* **24**, 54–69 (2011).
30. Paillard, D. Quaternary glaciations: from observations to theories. *Quat. Sci. Rev.* **107**, 11–24 (2015).

Acknowledgements We thank the IODP and ECORD (European Consortium for Ocean Research Drilling) for drilling the GBR, and the Bremen Core Repository for organizing the onshore sampling party. Financial support of this research was provided by the JSPS KAKENHI (grant numbers JP26247085, JP15KK0151, JP16H06309 and JP17H01168), the Australian Research Council (grant number DP1094001), ANZIC, NERC grant NE/H014136/1 and Institut Polytechnique de Bordeaux. This paper is a contribution to the INQUA commission on Coastal and Marine Processes and the PAGES PALSEA2 programme.

Reviewer information *Nature* thanks P. Whitehouse and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Y.Y. and J.M.W. were co-chief scientists of Expedition 325. J.O. and Y.Y. conducted GIA modelling. Y.Y. and T.M.E. wrote the manuscript in collaboration with J.M.W., A.L.T., J.-C.B. and M.H., and the paper was refined by contributions from the rest of the co-authors.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0335-4.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0335-4.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to Y.Y.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

IODP Expedition 325: the GBR environmental changes. The IODP Expedition 325 was designed to complement the previous IODP 'Tahiti Sea Level' Expedition 310 to Tahiti[31]. In preparation for Expedition 325, potential drill sites were surveyed with the CSIRO ship RV *Southern Surveyor* using multibeam sonar, seismics, an autonomous underwater vehicle and rock dredging[12,32]. At most of the GBR locations, the shelf breaks at approximately 120 m and is populated with prominent terrace-like structures; other relict reefs appear successively at depths[33–35] of 100–90 m, 60–50 m and 35–40 m.

The mission-specific platform chosen for Expedition 325 was the Greatship *Maya*, an IMO class II vessel capable of being positioned dynamically for geotechnical coring[32,35]. The expedition took place between 12 February and 6 April 2010. A total of 34 holes across 17 sites were sampled, ranging in depth from 46.4 m to 170.3 m such that the recovered coralgal deposits span several crucial but poorly defined periods during the LGM and last deglaciation. Sampling occurred at three locations along the northeastern coast of Australia, but this study focuses on one transect at 19.7° latitude offshore of Cairns at Noggin Pass (NOG-01B), and another at 17.1° latitude, offshore of Mackay at Hydrographer's Passage (Fig. 1, HYD-01C). Photographs of half-sectioned reef cores, relevant to the present study, are shown in Extended Data Figs. 1 and 2 together with the depths and genera of the dated corals and coralline algae. Classification of coral genera and species in terms of their habitat preferences according to depth and turbulence levels is also supported by considering the habitats of associated coralline algae taxa and crust thickness and vermetid gastropods. For example, individual coralline algae can have a limited range of habitats—bounded by sensitivity to light levels, wave energy and other factors—and can be used to more accurately constrain the palaeodepth ranges. Using this careful and detailed multi-proxy approach, each dated coral and coralline algae sample was placed within an internally consistent coralgal assemblage and palaeobathymetric scheme[36,37], enabling the construction of an independent RSL envelope at each site (Extended Data Figs. 1–3, Extended Data Table 1 and Supplementary Information).

Main lithologic facies observed in the Expedition 325 cores. The main lithologies are divided into coral reef framework and detrital sedimentary facies. The three boundstone facies are defined by their varying proportions of corals, coralline algae and microbial deposits forming coralgal, coralgal–microbialite and microbialite-dominated boundstones. The detrital facies can occur locally as internal sediments within the boundstones, or as metre-scale intervals of packstones to rudstones and unconsolidated sediments. Details of facies and depth estimates using facies as well as coral and coralline algae assembly can be found in Extended Data Figs. 1 and 2 and Extended Data Table 1, and are derived from ref. [12].

Reef 2 hiatus and GMSL drop to LGM-b. A major growth hiatus is observed in the inner shelf terrace at 104–106 m below sea level, at both HYD-01C (Hole M0031-33A) and NOG-01B (Hole M0055A) (Extended Data Figs. 3 and 4)[12]. This represents the turn-off (death) of Reef 2 about 21 kyr ago and is interpreted to be caused by the drop in sea-level to the LGM-b. The coralgal assemblages show that palaeowater depths were shallow (<10 m) just before the death of Reef 2, and lithologic and seismic evidence indicates that Reef 2 was a major subaerial exposure surface[12]. Furthermore, detailed scanning electron microscopic, energy-dispersive X-ray spectrum, X-ray diffraction analyses and thin-section observations of Reef 2 deposits confirm that they were exposed to freshwater or subaerial environments (for example, low-magnesium calcite cements in Hole M0055A Core 4R1) during the sea level lowstand at LGM-b (Extended Data Fig. 5). At this time shallow reef development migrated about 1–0.25 km seawards (that is, Reef 3a) in <2 kyr, as the RSL sea level fell to 118 m below present by about 20.5 kyr ago. The Reef 3a deposits and the older-than-MIS3 deposits are characterized by wholly marine diagenetic features consistent with the interpretation that the LGM-b sea level did not fall below this level (Extended Data Fig. 6). Sea level rose during the deglacial, causing major Reef 3a aggradation[12] before re-flooding the inner shelf terraces about 16.5 kyr ago and causing the re-establishment of the reef (Reef 3b) over its former position, marking the end of the hiatus at the top of Reef 2.

Determining the ages of sea level indicators. More than 165 representative coral skeletons and their aragonite content were analysed by powder X-ray diffraction, X-radiography, scanning electron spectroscopic and petrologic investigations, all of which confirmed the pristine nature of the dated samples[2]. In a few cases, X-ray diffraction picked up minor signatures of high-magnesium calcite, probably due to trace amounts of coralline algae and microbialite sediment. Yet no notable calcite peaks were observed in most cases. Physical cleaning of branched corals for U–Th dating and severe acid dissolution (more than 50% of their weight) of samples for radiocarbon dating was used to remove potential secondary precipitated materials. For massive *Porites* corals, physical cleaning is difficult, requiring further geochemical tests including inductively coupled plasma mass spectrometry analyses (ICP-MS). Skeletal Mg/Ca ratios confirmed the absence of large amounts of high-magnesium calcite and secondary aragonite cements. Even when secondary aragonite was found, the ages are not much affected because the form of the cements is indicative of the early post-mortem phase of corals. Further evaluations included limits on total uranium, ^{232}Th content and initial ^{234}U/^{238}U ratio. We applied different initial ^{234}U/^{238}U criteria: for samples of the deglacial period between 17 kyr ago and the present, the acceptable range was 1.1452 ± 0.0140, whereas for the samples from 30 kyr ago to 17 kyr ago 1.1402 ± 0.0140 was used. All data used to reconstruct the RSL envelopes for each transect are shown in Supplementary Information (see Methods section 'RSL reconstruction' for more details) and the primary samples used to determine the specific RSL inflection points are indicated as 'HY-1, 2' and 'NO-1, 2' for HYD-01C and NOG-01B, respectively, in Supplementary Information and highlighted in bold. U-series and radiocarbon ages from the same coral samples also showed remarkable consistency, along with radiocarbon ages on directly adjacent coralline algae. Taken together, and combined with the consistent reproducibility of the RSL envelopes between two transects, more than 500 km apart, this confirms the veracity of the data.

Radiocarbon dating. More than 500 radiocarbon dates were obtained using corals and coralline algae samples, which were all processed at the Atmosphere and Ocean Research Institute at the University of Tokyo to convert them into graphite[38]. Typically, 1 mg or more of graphite was measured using a Single Stage Accelerator Mass Spectrometer at the Atmosphere and Ocean Research Institute[39] and at the Australian National University (ANU)[40]. The results were then converted to calendar ages with local reservoir ages (12 ± 10 yr) from ref. [41], which we obtained by averaging between Heron Island (8 ± 6 yr) and Abraham Reef values (15 ± 6 yr). The calibration was then performed using international calibration datasets (IntCal 13 and Marine 13)[42].

Analytical procedures, mass spectrometry and U–Th dating. The analytical data are listed in Supplementary Information. Consistency between laboratories for replicate measurements of a single specimen is within 100 years, which is similar to the intra-coral variability observed in some specimens measured using the high-precision (Woods Hole Oceanographic Institute) method. Uranium series dating were conducted at three different laboratories: the Australian National University (ANU), the University of Oxford and the Woods Hole Oceanographic Institute. A 61-cm mass spectrometer is used at ANU, which can operate in charge mode[43]. The ^{229}Th, ^{230}Th and ^{232}Th isotopes were measured simultaneously in charge mode in Faraday cups using 20-pF feedback capacitors as active electrometer elements. The ^{233}U, ^{234}U and ^{235}U isotopes were also measured in charge mode, whereas ^{238}U was simultaneously measured using a 10^{10}-Ω feedback resistor. The magnitude of the ^{238}U low-mass tail was monitored continuously at mass ^{237}U in charge mode. This was used to subtract the ^{238}U tail from under the ^{233}U, ^{234}U and ^{235}U isotopes. Extensive measurements with an unspiked U-standard HU-1 showed that the shape of the ^{238}U tail remained invariant under a wide range of conditions, in particular, at the expected locations of the ^{233}U, ^{234}U, ^{235}U and ^{236}U peaks. Sample loads, on single rhenium filaments, ranged from 0.5 μg to 0.8 μg and the ^{238}U beam intensity was kept between 8×10^{11} A to 10×10^{11} A for several hours. At these intensities, the 10^{10}-Ω feedback resistor was used to avoid response-time problems encountered with the considerably slower 10^{11}-Ω resistors. The instrument was calibrated with reference to a secular-equilibrium standard HU-1.Comparisons with Western Australian last interglacial samples[44] and with Hulu Cave speleothem data[45] showed precise agreement with previous measurements. Sample processing followed previously established procedures[46]. U and Th were separated from the coral carbonate using U-Teva resin in a single pass.

U–Th dating at the University of Oxford and the Woods Hole Oceanographic Institute were measured by multi-collectors with ICP-MS. At the University of Oxford, U and Th isotopes were measured with ion counter collectors for the minor isotope beams. Approximately 0.3 g of coral sample was dissolved and spiked with a mixed ^{236}U:^{229}Th tracer. U and Th were purified and measured separately: U isotopes, statically; and Th by peak-hopping the 229 and 230 beams into an ion counter while normalizing beam intensity between the steps with either ^{232}Th or ^{235}U measured in Faraday collectors. Instrumental biases and relative collector efficiencies are accounted for using standard sample bracketing using U and Th isotope standards[47].

At the Woods Hole Oceanographic Institute, U and Th isotopes were measured by multi-collector ICP-MS in static mode with all isotopes in Faraday collectors[48]. Large (about 5 g) subsamples of coral were dissolved and spiked with a mixed ^{233}U:^{236}U:^{229}Th tracer, optimized for the LGM-to-deglacial age samples, and co-precipitated with Fe. To determine the ^{230}Th/^{238}U, purified U and Th fractions are recombined such that U and Th are measured together at isotope ratios that can be closely matched to bracketing standards. The ^{234}U/^{238}U is similarly determined statistically in Faraday collectors but on an unspiked aliquot.

All activity ratios and ages are calculated using the half-lives reported in ref. [45]. Ages are presented in Extended Data Table 1 as 'raw', assuming all ^{230}Th is accumulated in the coral since growth, and age-corrected for detrital ^{230}Th. The detrital correction makes use of the measured ^{232}Th/^{238}U as a proxy for the amount of detrital contamination, an assumed detrital composition of crustal origin[49], and an allowance for non-secular equilibrium of the contaminant.

© 2018 Springer Nature Limited. All rights reserved.

RSL reconstruction. The sample context was assessed using established criteria[12,32] including: (1) core quality, (2) orientation of well preserved corallites; (3) thick coralline algal crusts capping upper coral surfaces; (4) evidence of substrate attachment; and (5) the presence or absence and orientation of geopetals in lithified facies. Using these criteria, all the samples were classified into the following four context categories: (1) IS = in situ (convincing supporting evidence), (2) IS? = likely in situ (inclusive supporting evidence), (3) ISX = not in situ (convincing non-supporting evidence, and (4) ISN = status not known (inadequate evidence either way). Samples from highly drill-disturbed or poor recovery intervals were excluded. A total of 540 samples satisfying these criteria were used to construct a RSL envelope (upper and lower bounds) at both sites. Despite known temporal differences in ocean reservoir age, the coral U–Th and coral/coralline ^{14}C ccelerator mass spectrometry ages are remarkably consistent. However, wherever possible we used the more precise U–Th coral ages to constrain the upper and lower bounds of the envelopes. This was achieved by visually fitting a line through the dates that were >1 m apart and outside their analytical age errors, while also taking into account the upper bound of the palaeowater depth estimate of each sample and any core recovery uncertainties (Supplementary Information). Where multiple coral dates overlapped (in time) we used the mid-point between samples. If replicate age determinations were available for the same sample (that is, same coral colony or coralline algal crust) (Supplementary Information) an average age was calculated and plotted in Fig. 3. The upper bound or minimum position of the sea level envelope was further constrained by considering the overlapping palaeowater depth ranges of both the shallow-water sea level indicators and their coeval, deeper fore-reef slope equivalents. Finally, the major inflection points marking clear changes in the direction, amplitude and rate of RSL change were also identified for each envelope (Fig. 3). Thus the lower bound (that is, maximum sea level position) of RSL curves and the specific samples defining them are indicated in bold in Supplementary Information (that is, HY-1, 2 and NO-1, 2) and a close-up of the key samples constraining LGM-b is also shown in Extended Data Fig. 6.

GIA model predictions. The GIA model calculations included an Earth model describing the viscoelastic properties of the solid Earth, as well as an ice component documenting the ice melting history, reconstructed mainly from far-field sea level observations[4,50]. The Earth model is based on the seismologically derived Preliminary Reference Earth Model (PREM)[51] and consisted of an elastic lithosphere with an upper- and lower-mantle divide at 670-km depth. The lithosphere thickness was 70 km and upper- and lower-mantle viscosities ranged from $(1–10) \times 10^{20}$ Pa s and $(1–100) \times 10^{22}$ Pa s respectively. This model provides an accurate treatment of time-dependent continental shorelines[52] and the Earth rotation feedback on sea level[53].

The ice model described above was adjusted to match the newly obtained (RSL) records from the GBR. The analytical uncertainties in RSL were taken into account and the shallow and deep extremes of the RSL envelope were tested. We first employed the ice history model developed by the ANU group[4] and ran the GIA model to obtain RSL histories for the GBR. The ANU ice model with the same relative ice volumes was then scaled to fit the NOG and HYD RSL. The scaling was done manually within a reasonable range of various parameters and by keeping the relative ice volume of various ice sheets the same as in the original ANU model, though keeping the Eurasian ice model almost the same as the ANU model since the history of this ice sheet is reasonably well constrained from both observations and models[54–56] compared to other ice sheets. The chosen Earth parameters (lithospheric thickness 70 km, upper-mantle viscosity 2×10^{20} Pa s, and lower-mantle viscosity 10^{22} Pa s)[57,58] fit the GBR region Holocene sea levels well. The analytical uncertainties in RSL and the range in Earth model parameters (approximately ±2.5 m contribution to GMSL) were used in calculating the MAX and MIN extremes (Extended Data Figs. 8 and 9). Supplementary Information shows GMSL contributions individually for each major ice sheet and for the ANU and highest and lowest sea level GMSL scenarios (in eustatic terms). The two GMSL curves were then used to construct RSLs in far-field sites with previously published RSL data for comparison (Fig. 4 and Extended Data Fig. 7). Potential Earth model uncertainties were also considered with variable lithosphere thickness and a range of viscosities, of the lower as well as the upper mantle, resulting in more than 60 GIA model experiments. Shaded areas of curves represent possible ranges in RSL for individual sites (Fig. 4 and Extended Data Fig. 7). Visual inspection of the results indicates that the MAX model fits the data remarkably well for almost all the far-field locations tested. In turn, this indicates that the shallow coral habitat depth estimates appear to be sufficiently robust without necessitating extended, deeper-water limits.

During the LGM (30–19 kyr ago; Fig. 3), water depth uncertainties for most samples from the HYD and NOG transects are <5 m. Figure 2 shows RSL curves derived from MAX and MIN models. Various Earth rheology parameters were also tested using both MAX and MIN ice models where the shaded region around GMSLs in Extended Data Fig. 9 represents the corresponding range in RSLs. During the LGM, the maximum magnitude of RSL difference[59] between the two transects is less than 10 m and hence RSL variations arising from hydro-isostasy are small. The models were run over the maximum possible ranges of the rheological parameters so that the range of RSLs depicted in the shaded zone in Extended Data Fig. 9 cover the full range of possibilities. Traditionally, lower-mantle viscosity has been estimated as about 10^{22} Pa s using far-field RSL observations[57,58], whereas recent studies have reported much higher values[4] of about 7×10^{22} Pa s. Thus, assuming a typical lithospheric thickness and upper-mantle viscosity, respectively, of 70 km and 2×10^{20} Pa s, the maximum RSL differences associated with the above range of lower-mantle viscosities is about 5 m (Extended Data Fig. 9). This number is smaller than the typical uncertainties inherent in RSL observations at the GBR and is therefore well suited for reconstructions of GMSLs during LGM-a and LGM-b. In summary, we concluded that the MAX model provides the best estimate of GMSLs as well as indicating that these tighter depth uncertainties for GBR corals[12] do provide consistent results. Therefore, the extended MIN to MAX range, employed here, covers the likely range of GMSL constructions with confidence, as can be ascertained by visual inspection (Extended Data Fig. 9).

Glaciological evidence, including from ice cores, cannot easily accommodate the required increase in ice volume. However, the total increase is shared among the large ice sheets (Extended Data Fig. 8). Ice cores retrieved from Antarctica and Greenland are not able to resolve the required magnitude of elevation changes in continental interiors. It is also likely that current ice-free regions may have been places that retained the extra ice at these times. For example, recent evidence suggests that an extensive ice sheet was grounded on the Ross Sea for at least 3,700 years[60]. New bathymetric data as well as glacial models support these conclusions[61]. However, there is still scope to improve the glaciological models and we hope that our data will contribute to this effort.

Discrepant GMSLs during the LGM at either −120 m or −140 m have been reported respectively for Barbados[7] and the Bonaparte Gulf in North Australia[5]. This has now been reconciled using the recently reported Earth rheology model with J2 observations[10] as well as considering subducting material in Barbados[62]. The model included 65–100 km of lithospheric thickness and upper- and lower-mantle viscosities of $(1–3) \times 10^{20}$ Pa s and 10^{23} Pa s. The global RSL observations could reasonably be explained if GMSL during the LGM were about −130 m. This finding is consistent with the model derived from more than 1,000 far-field RSL observations[4]. The results from the present study, for both the MAX and the MIN options, are respectively −125 m and −130 m and thus consistent with the independent estimates described above.

Data availability. All the data used in this manuscript are available in Supplementary Information. The modified ANU ice model is available upon reasonable request.

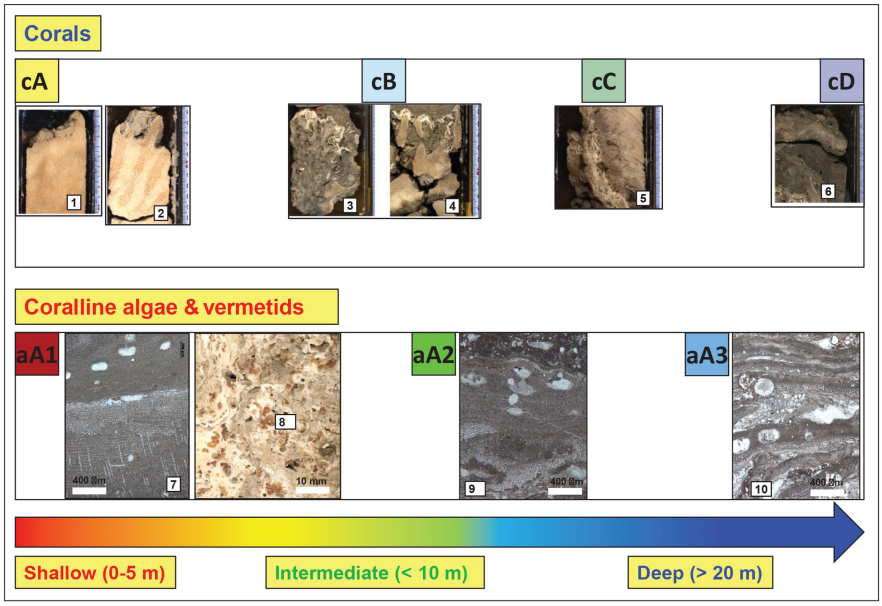
Code availability. The model that we employed in this paper is available from J.O. (okuno@nipr.ac.jp) and Y.Y. upon reasonable request.

31. Camoin, G. F. et al. Reef response to sea-level and environmental changes during the last deglaciation. IODP Expedition 310 "Tahiti Sea Level". *Geology* **40**, 643–646 (2012).
32. Webster, J. M., Yokoyama, Y., Cotterill, C. & the Expedition 325 Scientists. *Great Barrier Reef Environmental Changes*. http://publications.iodp.org/proceedings/325/325title.htm (Proc IODP 325, Integrated Ocean Drilling Program, 2011).
33. Abbey, E., Webster, J. M. & Beaman, R. J. Geomorphology of submerged reefs on the shelf edge of the Great Barrier Reef: the influence of oscillating Pleistocene sea-levels. *Mar. Geol.* **288**, 61–78 (2011).
34. Bridge, T. C. L. et al. Topography, substratum and benthic macrofaunal relationships on a tropical mesophotic shelf margin, central Great Barrier Reef, Australia. *Coral Reefs* **30**, 143–153 (2011).
35. Yokoyama, Y. et al. IODP Expedition 325: Great Barrier Reefs Reveals Past Sea-Level, Climate and Environmental Changes since the Last Ice Age. *Sci. Drill.* **12**, 32–45 (2011).
36. Cabioch, G., Montaggioni, L. F., Faure, G. & Ribaud-Laurenti, A. Reef coralgal assemblages as recorders of paleobathymetry and sea-level changes in the Indo-Pacific province. *Quat. Sci. Rev.* **18**, 1681–1695 (1999).
37. Dechnik, B. et al. The evolution of the Great Barrier Reef during the Last Interglacial Period. *Global Planet. Change* **149**, 53–71 (2017).
38. Yokoyama, Y., Miyairi, Y., Matsuzaki, H. & Tsunomori, F. Relation between acid dissolution time in the vacuum test tube and time required for graphitization for AMS target preparation. *Nucl. Instrum. Methods B* **259**, 330–334 (2007).
39. Hirabayashi, S., Yokoyama, Y., Suzuki, A., Miyairi, Y. & Aze, T. Multidecadal oceanographic changes in the western Pacific detected through high-resolution bomb-derived radiocarbon measurements on corals. *Geochem. Geophys. Geosyst.* **18**, 1608–1617 (2017).
40. Fallon, S. J., Fifield, L. K. & Chappell, J. M. The next chapter in radiocarbon dating at the Australian National University: status report on the single stage AMS. *Nucl. Instrum. Methods B* **268**, 898–901 (2010).
41. Druffel, E. R. M. & Griffin, S. Variability of surface ocean radiocarbon and stable isotopes in the southwestern Pacific. *J. Geophys. Res. Oceans* **104**, 23607–23613 (1999).

© 2018 Springer Nature Limited. All rights reserved.

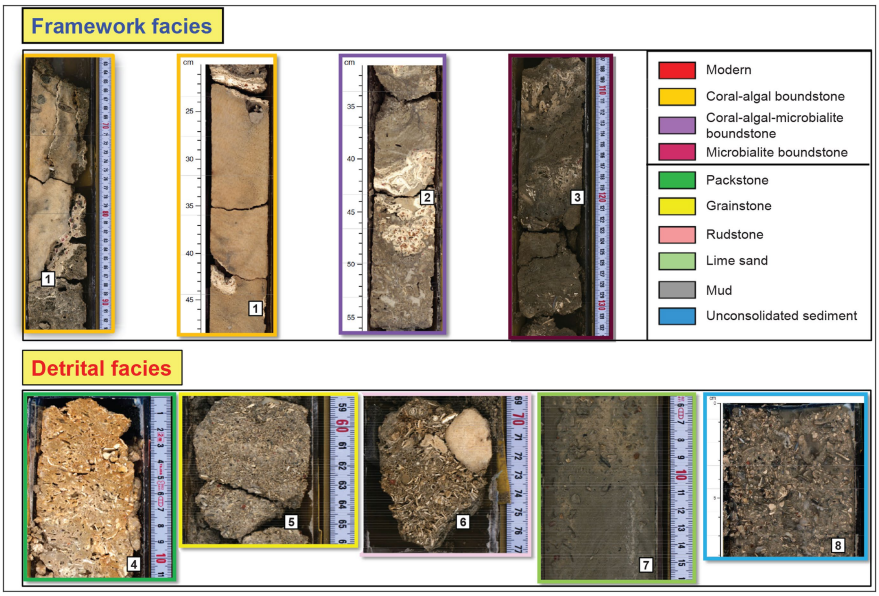
42. Reimer, P. J. et al. INTCAL13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal BP. *Radiocarbon* **55**, 1869–1887 (2013).
43. Esat, T. M. Charge collection thermal ion mass spectrometry of thorium. *Int. J. Mass Spectrom. Ion Process.* **148**, 159–170 (1995).
44. Stirling, C. H., Esat, T. M., Lambeck, K. & McCulloch, M. T. Timing and duration of the last interglacial; evidence for a restricted interval of widespread coral reef growth. *Earth Planet. Sci. Lett.* **160**, 745–762 (1998).
45. Cheng, H. et al. The half-lives of uranium-234 and thorium-230. *Chem. Geol.* **169**, 17–33 (2000).
46. Stirling, C. H., Esat, T. M., McCulloch, M. T. & Lambeck, K. High-precision U-series dating of corals from Western Australia and implications for the timing and duration of the Last Interglacial. *Earth Planet. Sci. Lett.* **135**, 115–130 (1995).
47. Thomas, A. L. et al. Penultimate deglacial sea level timing from uranium/thorium dating of Tahitian corals. *Science* **324**, 1186–1189 (2009).
48. O'Leary, M. J. et al. Ice sheet collapse following a prolonged period of stable sea level during the last interglacial. *Nat. Geosci.* **6**, 796–800 (2013).
49. Taylor, S. R. & McLennan, S. M. *The Continental Crust: Its Composition and Evolution: An Examination of the Geochemical Record Preserved in Sedimentary Rocks* (Blackwell Scientific, Oxford, 1985).
50. Okuno, J., Nakada, M., Ishii, M. & Miura, H. Vertical tectonic crustal movements along the Japanese coastlines inferred from late Quaternary and recent relative sea-level changes. *Quat. Sci. Rev.* **91**, 42–61 (2014).
51. Dziewonski, A. M. & Anderson, D. L. Preliminary reference Earth model (PREM). *Phys. Earth Planet. Inter.* **25**, 297–356 (1981).
52. Lambeck, K., Purcell, A., Johnston, P., Nakada, M. & Yokoyama, Y. Water-load definition in the glacio-hydro-isostatic sea-level equation. *Quat. Sci. Rev.* **22**, 309–318 (2003).
53. Boulton, G. S., Dongelmans, P., Punkari, M. & Broadgate, M. Palaeoglaciology of an ice sheet through a glacial cycle: the European ice sheet through the Weichselian. *Quat. Sci. Rev.* **20**, 591–625 (2001).
54. Lambeck, K., Smither, C. & Johnston, P. Sea-level change, glacial rebound and mantle viscosity for northern Europe. *Geophys. J. Int.* **134**, 102–144 (1998).
55. Lambeck, K., Purcell, A., Zhao, J. & Svensson, N.-O. The Scandinavian ice sheet: from MIS 4 to the end of the Last Glacial Maximum. *Boreas* **39**, 410–435 (2010).
56. Milne, G. A. & Mitrovica, J. X. Post glacial sea-level change on a rotating earth. *Geophys. J. Int.* **133**, 1–19 (1998).
57. Nakada, M. & Lambeck, K. Late Pleistocene and Holocene sea-level change in the Australian region and mantle rheology. *Geophys. J.* **96**, 497–517 (1989).
58. Lambeck, K. & Nakada, M. Late Pleistocene and Holocene sea-level change along the Australian coast. *Global Planet. Change* **3**, 143–176 (1990).
59. Yokoyama, Y., Purcell, A., Marshall, J. F. & Lambeck, K. Sea-level during the early deglaciation period in the Great Barrier Reef, Australia. *Global Planet. Change* **53**, 147–153 (2006).
60. Bart, P., Krogmeier, B. J., Bart, M. P. & Tulaczyk, S. The paradox of a long grounding during West Antarctic Ice Sheet retreat in Ross Sea. *Sci. Rep.* **7**, 1262 (2017).
61. Halberstadt, A. R. W., Simkins, L. M., Greenwood, S. L. & Anderson, J. B. Past ice-sheet behaviours: retreat scenarios and changing controls in the Ross Sea Antarctica. *Cryosphere* **10**, 1003–1020 (2016).
62. Austermann, J., Mitrovica, J. X., Latychev, K. & Milne, G. A. Barbados-based estimate of ice volume at Last Glacial Maximum affected by subducted plate. *Nat. Geosci.* **6**, 553–557 (2013).
63. Edwards, R. L. et al. A large drop in atmospheric $^{14}C/^{12}C$ and reduced melting in the Younger Dryas, documented with ^{230}Th ages of corals. *Science* **260**, 962–968 (1993).
64. Tarasov, L. & Peltier, W. R. Coevolution of continental ice cover and permafrost extent over the last glacial-interglacial cycle in North America. *J. Geophys. Res.* **112**, F02S08 (2007).

© 2018 Springer Nature Limited. All rights reserved.



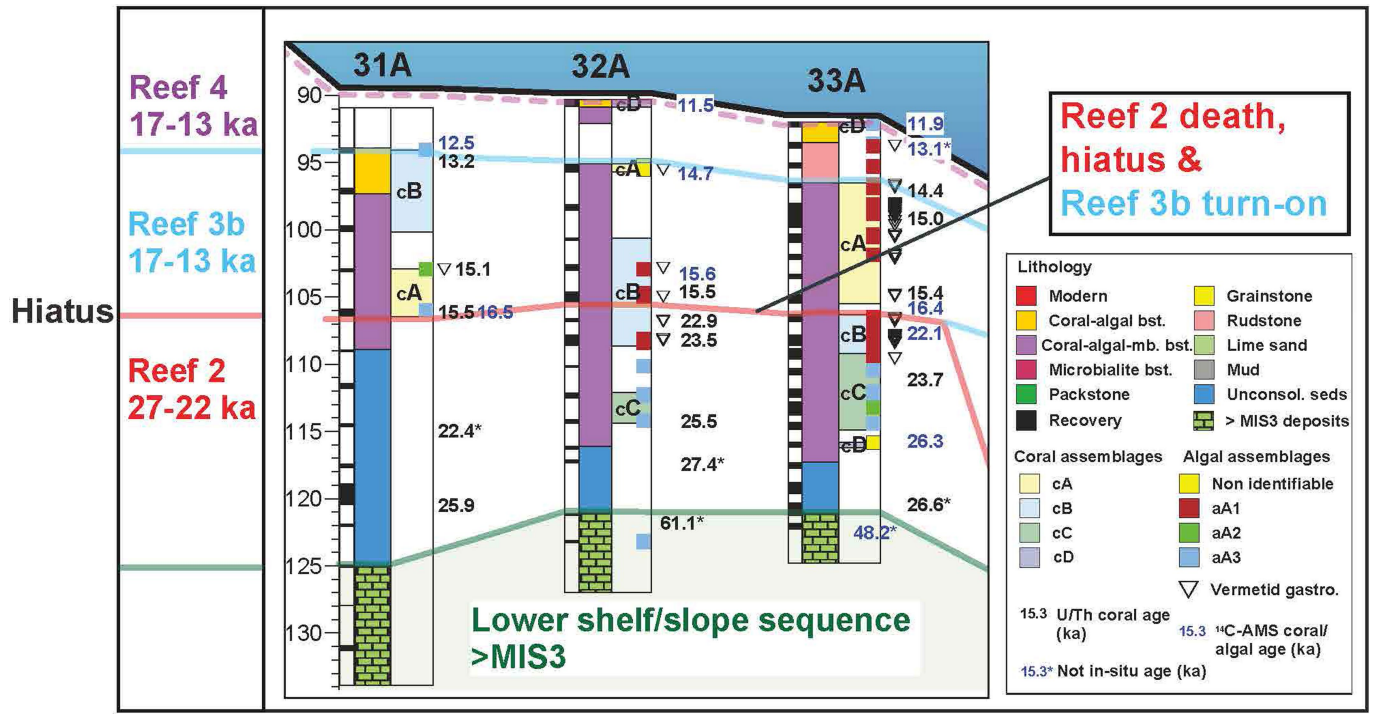
Extended Data Fig. 1 | Simplified classification of the main coralgal assemblages observed in the Expedition 325 cores. Shallow reef habitats are represented by coral assemblages when associated with thick crusts of aA1 coralline algae and vermetid gastropods (8). Deep, fore-reef slope settings are defined by coral assemblages cD when associated with thin crusts of aA3 and in the absence of aA1 and aA2. cA, massive/robust branching *Isopora* (1) and corymbose *Acropora* gr. *humilis* (2); cB, branching *Seriatopora* (3) and *Acropora* sp. (4); cC, massive/encrusting meruliniids (5); cD, encrusting to massive *Porites* (6) and encrusting *Montipora*; aA3, *Mesophyllum* and *Lithothamnion* (10); aA1, *Porolithon onkodes* (7); and aA2, thin *P. onkodes*, *Porolithon gardineri, Harveylithon gr. munitum* (9).

© 2018 Springer Nature Limited. All rights reserved.



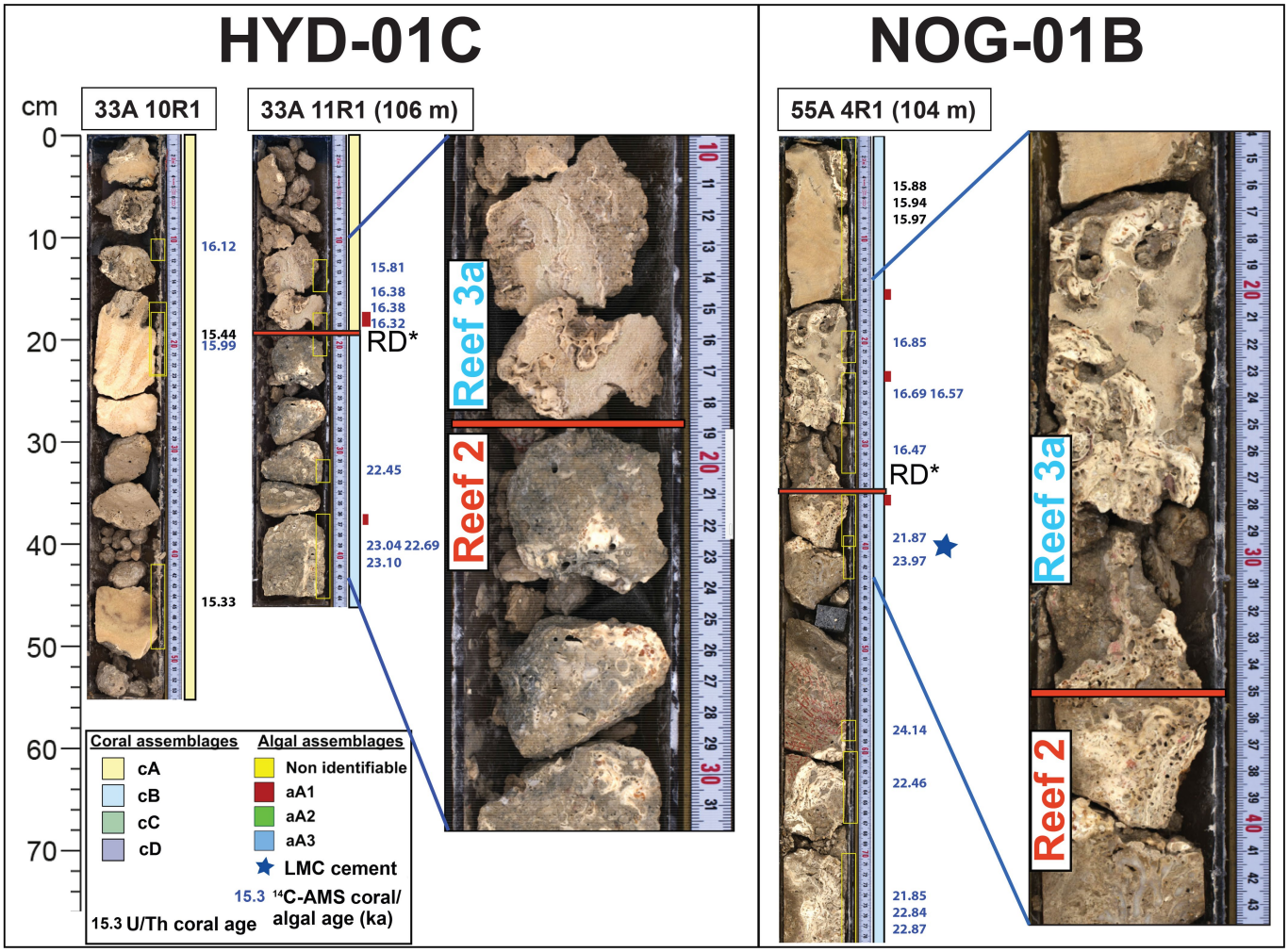
Extended Data Fig. 2 | Representative facies observed in the Expedition 325 cores. The main lithologies are divided into coral reef framework, (1)–(3), and detrital sedimentary facies, (4)–(8).

© 2018 Springer Nature Limited. All rights reserved.



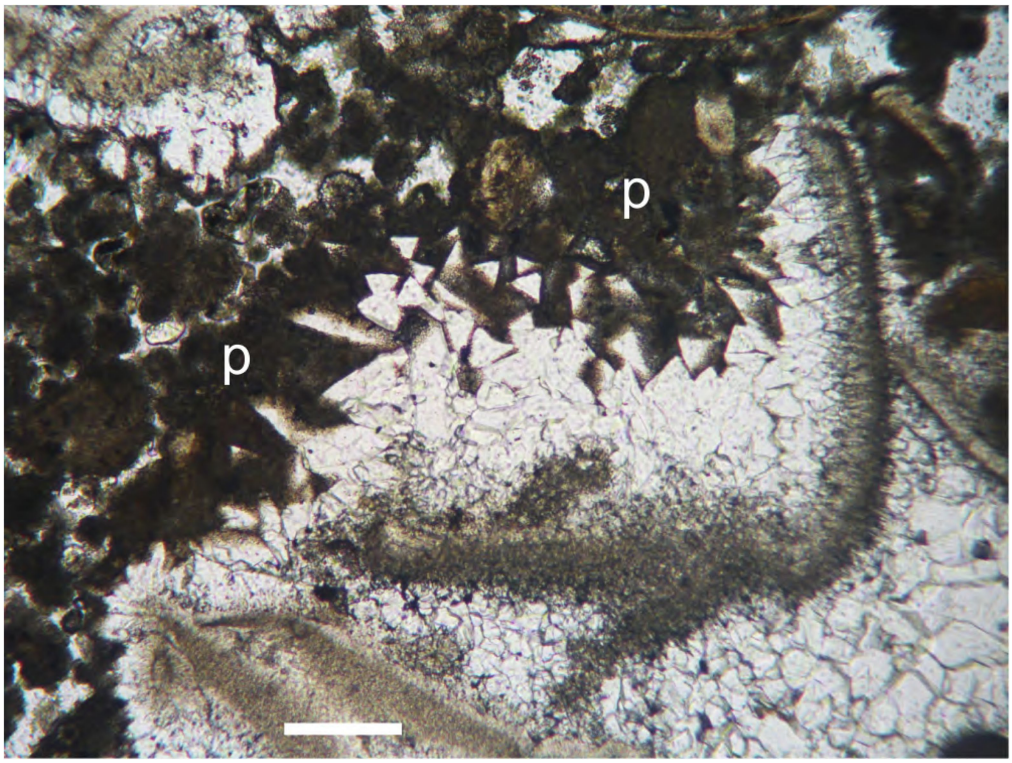
Extended Data Fig. 3 | Stratigraphic synthesis of inner terrace cores at Mackay. The vertical patterns in coral recovery, lithologies, coralgal assemblages and vermetid gastropods are summarized here. These data define a major hiatus (pink line) in reef development as the top of Reef 2 was exposed following the sea level fall to the LGM-b low-stand (see main text for details). Following deglacial sea level rise, reef growth was re-established as Reef 3b turned on when the shelf reflooded before this reef drowned after about 14 kyr ago (ka). ('bst.', boundstone; 'mb', microbialite; 'unconsol.', unconsolidated sediment; and 'gastro', gastropod.)

© 2018 Springer Nature Limited. All rights reserved.



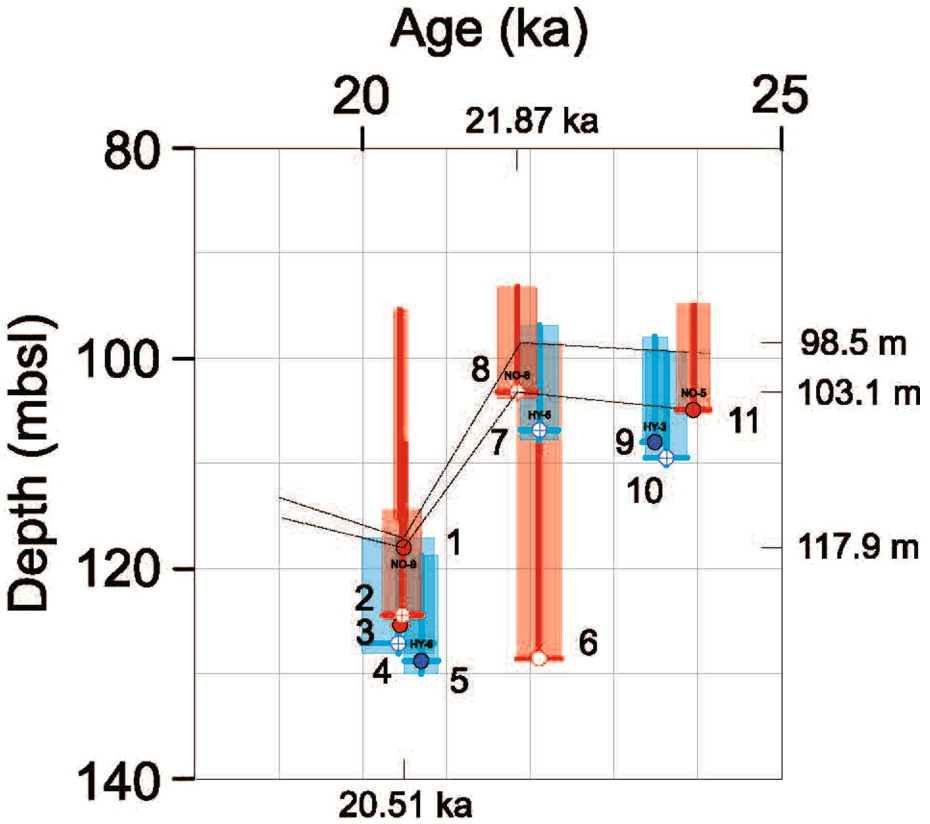
Extended Data Fig. 4 | Close-up core images showing the boundary between Reef 2 and Reef 3a. Cores M00033A-10R and 11R (HYD-01C) and M00055A-4R (NOG-1B) clearly define the nature of the Reef 2/ Reef 3b boundary, which represents a major hiatus in reef growth (see Extended Data Fig. 3). This boundary is characterized by major changes in lithologies, coralgal assemblages and diagenetic features, including fresh water and meteoric cements (blue star) indicating that the top of Reef 2 has been subaerially exposed. (LMC, low-magnesium calcite; RD, reef-death events.)

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 5 | Evidence of subaerial exposure. The blocky low-magnesium calcite meteoric cement that is related to subaerial exposure is initially precipitated in the intergranular voids of grainstone in 55A 5R1. Then peloids (p) of high-magnesium calcite formed under marine conditions and filled the remaining voids when they were submerged following re-flooding. Scale bar, 100 μm.

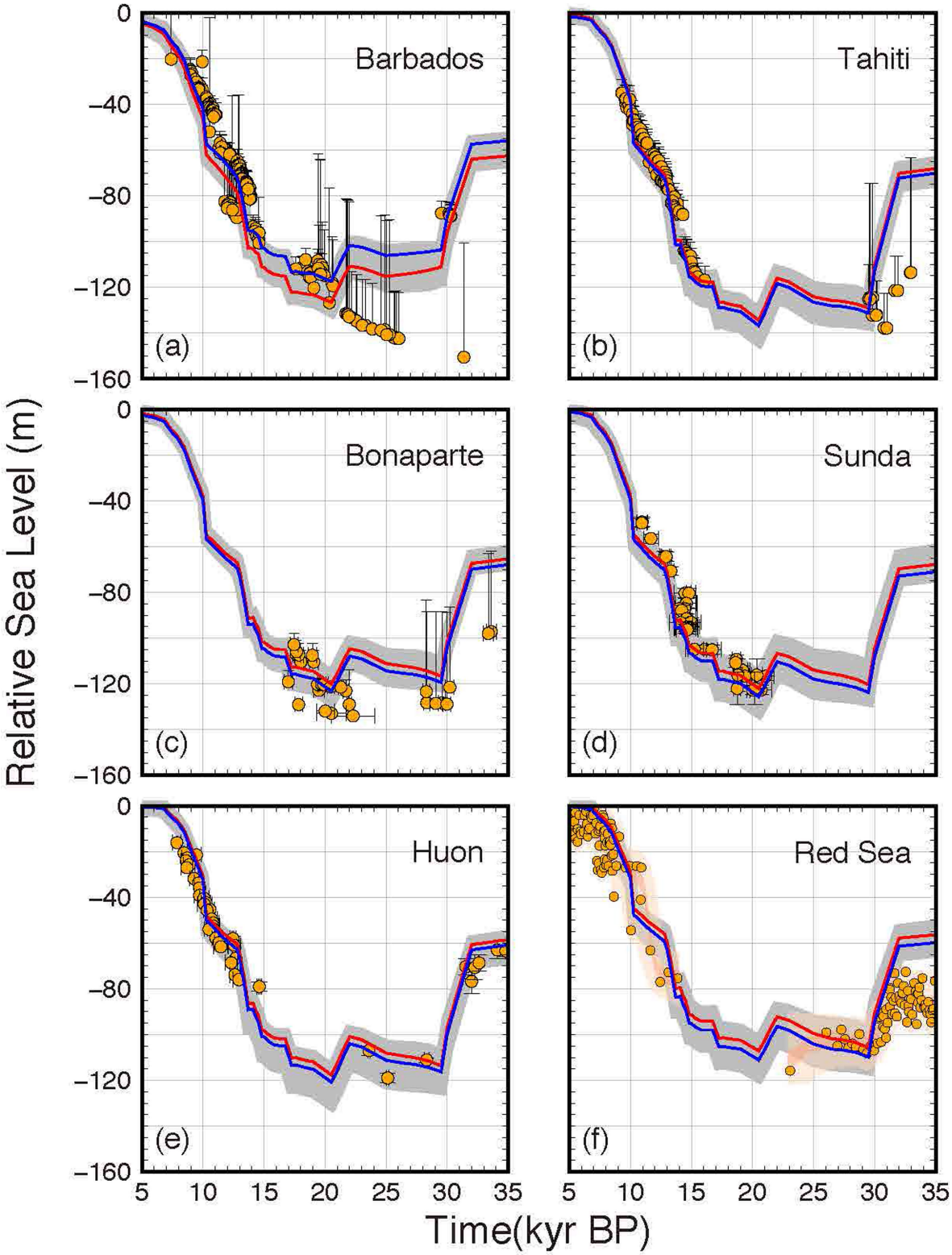
© 2018 Springer Nature Limited. All rights reserved.



	HYD-01C	NOG-01B
$^{14}C_{[corals]}$	○	○
$^{14}C_{[algae]}$	⊕	⊕
U/Th	●	●

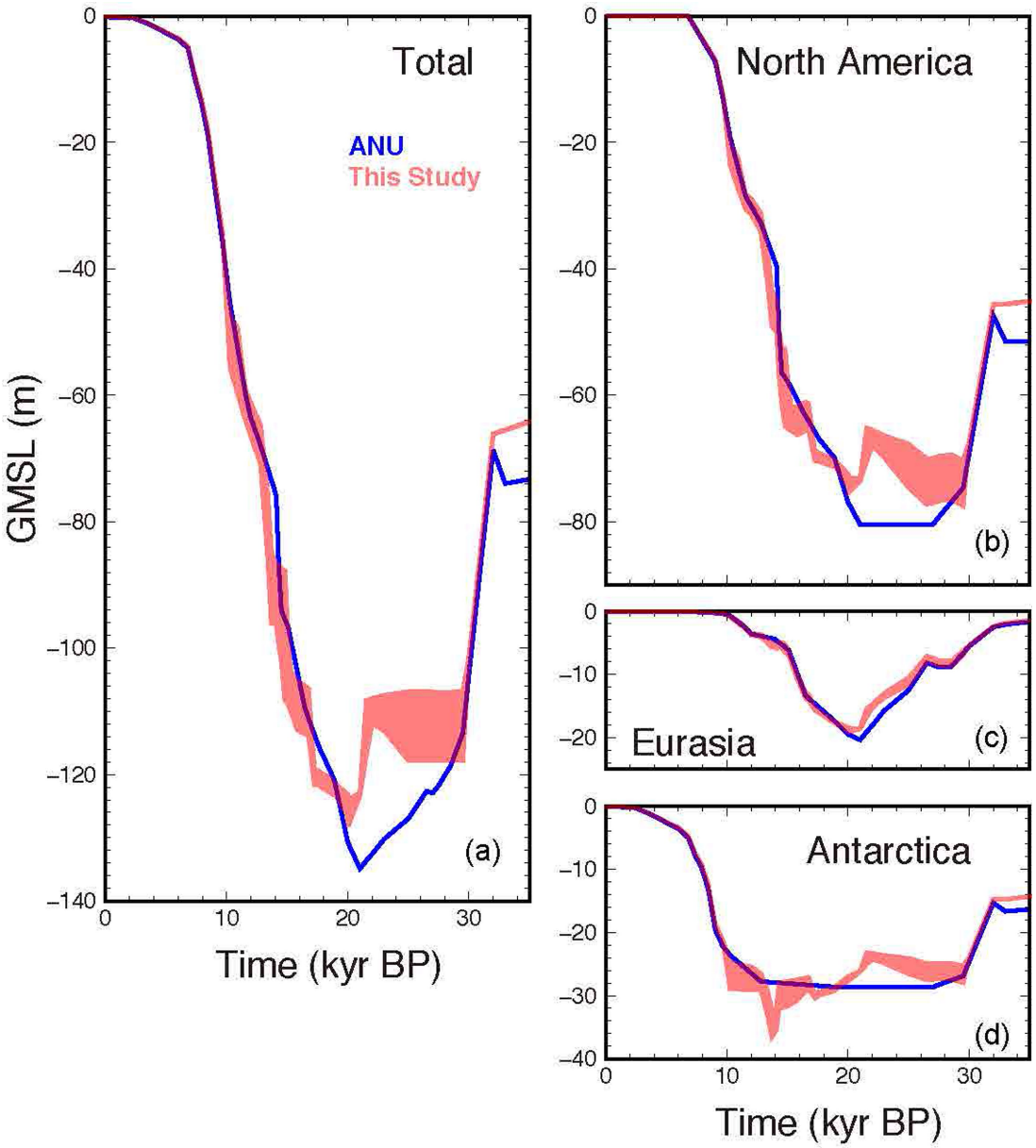
Extended Data Fig. 6 | Timing and extent of the sea-level drop at LGM-b. Age versus depth plot showing the key in situ RSL data points from Hydrographer's Passage (HYD-01C) (in blue) and Noggin Pass (NOG-01B) (in red). Accelerator mass spectrometer ^{14}C ages derived from corals are indicated by open circles and those derived from coralline algae are indicated by crossed circles. U–Th coral ages are indicated by filled circles. Inflection points defining the maximum position of the RSL at HYD-01C and NOG-01B are also shown (see labels NO-5, 8, 9, and HY-3, 5, 6 in Supplementary Information, corresponding respectively to data points 11, 8, 1 and 9, 7, 5 on this figure; see also Fig. 3). The combined RSL envelope represented by the black lines (maximum and minimum positions) takes into account the uncertainties in the age (2σ), palaeowater depth and position in the core of each data point, which is illustrated by a coloured rectangle (blue for Hydrographer's Passage and red for Noggin Pass) (see Methods for more details). If we were to omit the ^{14}C data owing to possible unaccounted-for variability in local reservoir ages[63], the LGM-b sea level drop defined by coral U–Th ages would be 1.5 kyr earlier at 23.5 ka, corresponding to an extended sea level drop over about 3 kyr at a rate of about 7 m kyr^{-1}. Key RSL index points are as follows. 1, 325-M0053A-13R-1W 21-25 (20.51 ka, 117.93 m; NO-9). 2, 325-M0054B-06R-1W 64-67 (20.50 ka, 124.39 m). 3, 325-M0054B-07R-1W 5-9 (20.47 ka, 125.3 m). 4, 325-M0035A-18R-1W 10-15 (20.43 ka, 127.11 m). 5, 325-M0036A-18R-2W 8-10 (20.70 ka, 128.75 m; HY-6). 6, 325-M0054B-08R-2W 73-75 (22.13 ka, 128.54 m). 7, 325-M0033A-11R-CCW 5-11 (22.11 ka, 106.83 m; HY-5). 8, 325-M0055A-04R-1W 35-40b (21.87 ka, 103.13 m; NO-8). 9, 325-M0032A-10R-1W 18-20 (23.49 ka, 107.95 m, HY-3). 10, 325-M0033A-13R-CCW 1-3 (23.62 ka, 109.46 m). 11, 325-M0055A-04R-2W 99-105 (23.97 ka, 104.84 m; NO-5).

© 2018 Springer Nature Limited. All rights reserved.



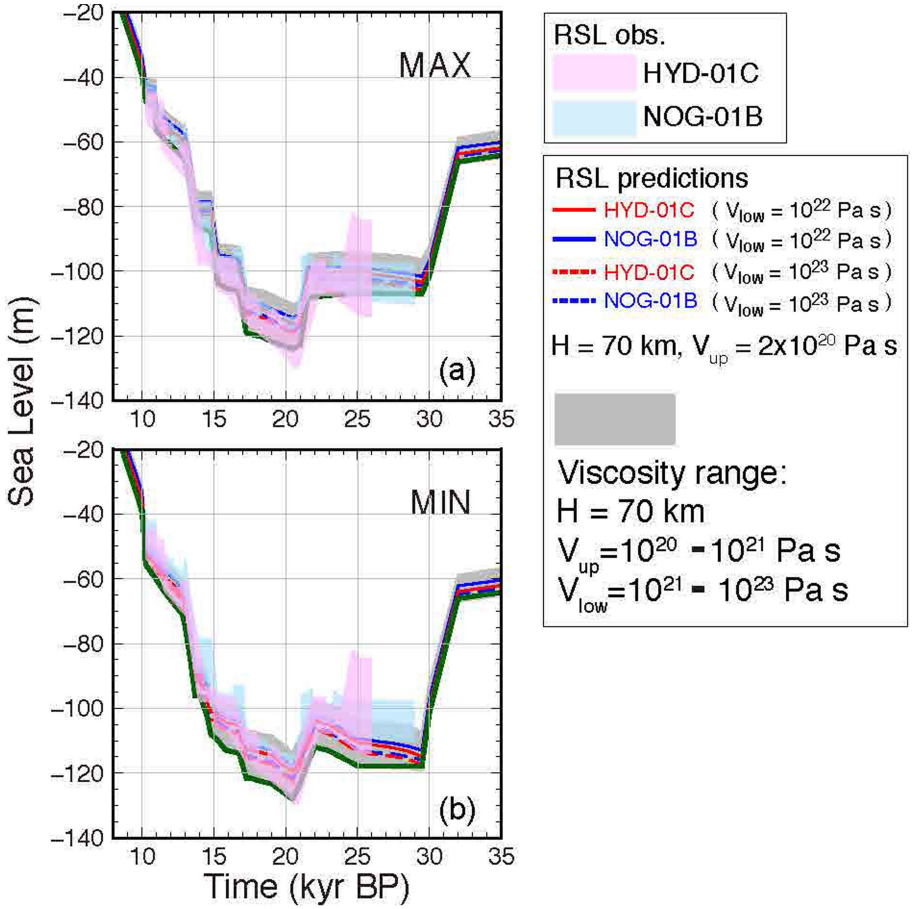
Extended Data Fig. 7 | GIA results for selected sites. Calculations, for selected far-field sites (**a**–**f**), using the ice model based on the lower-bound of RSL from GBR. GIA calculations implemented using parameters of upper-mantle viscosity and lithospheric thickness of 10^{20} Pa s and 70 km and lower-mantle viscosity of 10^{22} Pa s. However, the differences between data and calculations for the far-field sites of Tahiti (**b**), Bonaparte (**c**) and Sunda (**d**) indicate a better match with GIA when the higher RSL obtained from this study is used (Fig. 4) in calculating the global deglacial sea levels. The grey band represents the range of RSL predictions using GIA modelling with various earth parameters (lithospheric thickness $H = 70$ km, upper-mantle viscosity 10^{20}–10^{21} Pa s, lower-mantle viscosity 10^{21}–10^{23} Pa s), and a melting model (of ice history), which, in this case, was the MIN model. The red lines are for $H = 70$km, upper-mantle viscosity 2×10^{20} Pa s and lower-mantle viscosity 10^{22} Pa s. The blue lines show the case for $H = 70$ km, upper-mantle viscosity 2×10^{20} Pa s and lower-mantle viscosity 10^{23} Pa s. Vertical uncertainties are depth ranges for each sea level indicators and horizontal error bars are 2σ uncertainties for age estimations reported from the literature.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 8 | Comparisons between the ANU GMSL and the present GBR-based GMSL. a–d, Total (**a**), North America (**b**), Eurasia (**c**) and Antarctic (**d**) ice sheet growth and melting histories for the past 35 kyr inferred from the present study (red bands) are compared with previously reported model (blue line). The blue curves represent the ANU GMSL results[4], whereas the red bands, covering a range of GBR-RSL-based GMSL, are from this work. During the transition from LGM-a to LGM-b, around 21 kyr ago, there is enhanced precipitation over the North American Ice Sheet and to a lesser extent over Antarctica, although, for the latter, the ice volume continues to build up, for longer, until the termination of LGM-b approximately 17 kyr ago. The manually adjusted nominal ANU ice model will influence our inferred melting history for each ice sheet. Furthermore, although the ANU model differs in some respects from other ice sheet reconstructions[64], the results from these simulations are not very different.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | GIA model results for HYD and NOG with different viscosity settings. MAX (**a**) and MIN (**b**) represent the maximum and minimum extremes of GBR RSL. Blue and red bands are RSL ranges derived from our study for Hydrographer's Passage (HYD) and Noggin Pass (NOG). Grey bands represent the range of predicted sea levels using the new ice model. Lithospheric thickness is fixed at 70 km, whereas mantle viscosities for the upper and lower mantle varied between 10^{20}–10^{21} Pa s and 10^{21}–10^{23} Pa s. RSL predictions for representative viscosities for HYD-01C and NOG-01B are shown as red and blue solid and dotted lines respectively. V_{up} and V_{low} represent the upper- and lower-mantle viscosity values. (obs., observations.)

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Simplified coral and coralline algal assemblages and their likely palaeoenvironmental setting

Assemblage	Main components	Paleoenvironmental setting	Paleowater depth***
Corals			
cA	Massive/robust branching *Isopora* & medium corymbose branching *Acropora humilis* gr.	Shallow, high energy/exposed reef edge	0-10 m; **0-5 m***
cB	Fine branching *Seriatopora* and *Acropora* sp.	Wide range of protected reef settings	0-20 m; semi exposed, **0-10 m***
cC	Massive & encrusting meruliniids (*Dipsastrea, Cyphastrea, Platygyra*)	Wide range of protected to turbid reef settings	0-30 m; **0-10 m***
cD	Encrusting to Massive *Porites* & encrusting *Montipora*	Very wide range of reef settings	0-60 m; **0-10 m*, 20-102 m****
Coralline algae			
aA1	Thick crust of *Porolithon onkodes*	Shallow reef settings	**0-10 m**
aA2	Thin crusts (<0.2 mm) of *P. onkodes*. Potential occurrence of*Porolithon gardineri, Harveylithon gr. munitum*	Wide range of reef settings	**0-20 m**
aA3	Absence of aA1 & aA2. Occurrence of thin crusts of *Mesophyllum, Lithothamnion* and *Lithoporella*	Deep fore-reef slope settings	**>20 m**

*When associated with centimetres-thick aA1 CAR crusts and vermetid gastropods.
**When associated with thin crusts of aA3 and lacking vermetid gastropods.
***Palaeowater depths were estimated by comparison with their modern Indo-Pacific counterparts.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0299-4

Hominin occupation of the Chinese Loess Plateau since about 2.1 million years ago

Zhaoyu Zhu[1,2]*, Robin Dennell[3]*, Weiwen Huang[2,4], Yi Wu[5], Shifan Qiu[6], Shixia Yang[4,7], Zhiguo Rao[8], Yamei Hou[2,4], Jiubing Xie[9], Jiangwei Han[10] & Tingping Ouyang[1,11]

Considerable attention has been paid to dating the earliest appearance of hominins outside Africa. The earliest skeletal and artefactual evidence for the genus *Homo* in Asia currently comes from Dmanisi, Georgia, and is dated to approximately 1.77–1.85 million years ago (Ma)[1]. Two incisors that may belong to *Homo erectus* come from Yuanmou, south China, and are dated to 1.7 Ma[2]; the next-oldest evidence is an *H. erectus* cranium from Lantian (Gongwangling)—which has recently been dated to 1.63 Ma[3]—and the earliest hominin fossils from the Sangiran dome in Java, which are dated to about 1.5–1.6 Ma[4]. Artefacts from Majuangou III[5] and Shangshazui[6] in the Nihewan basin, north China, have also been dated to 1.6–1.7 Ma. Here we report an Early Pleistocene and largely continuous artefact sequence from Shangchen, which is a newly discovered Palaeolithic locality of the southern Chinese Loess Plateau, near Gongwangling in Lantian county. The site contains 17 artefact layers that extend from palaeosol S15—dated to approximately 1.26 Ma—to loess L28, which we date to about 2.12 Ma. This discovery implies that hominins left Africa earlier than indicated by the evidence from Dmanisi.

The Lantian area of the Shaanxi province lies on the southern margin of the Chinese Loess Plateau (Fig. 1), which covers about 270,000 square kilometres[7] (Fig. 1). Loess deposition in China since 2.6 Ma has been massive, with up to 100–300 m being deposited on the Loess Plateau[7]. Typical aeolian loess–palaeosol sequences are divided into 33 couplets of alternating palaeosol (S) and loess (L) sediments from top to bottom, from the S0–L1 couplet to the S32–L33 couplet[7,8]. This sequence records East Asian monsoon development since 2.6 Ma and correlates with marine isotope stage 1 to marine isotope stage 102 or 104[9,10]. Loess units correspond to even-numbered marine isotope stages (cold periods or glacial stages) and palaeosol units correspond to odd-numbered marine isotope stages (warm periods or interglacial stages). Numerous sections indicate that Loess Plateau sequences can be correlated over distances of >1,100 km from west to east and >600 km from north to south. Palaeomagnetic dating, which establishes palaeomagnetic reversal boundaries, is the most reliable method for dating these sequences; for example, the Matuyama/Brunhes boundary occurs in the S7–L8 couplet (and sometimes in S8); the Jaramillo subchron between L10 and S12; and the Olduvai subchron in the lower part of L25 to the base of S26[7,8,10–16] (see Supplementary Information). Through palaeomagnetic dating and astronomical tuning[10,14], the age of each loess and palaeosol has been calculated in two schemes to within a precessional cycle (Supplementary Table 1) and used in international correlations of the Quaternary geological timescale[17,18]; here we use the 'Chiloparts' timescale[10] (Supplementary Table 1). Because of questions about the positions of geomagnetic reversals in loess–palaeosol couplets[16,19] and time delays of about 10–30 thousand years for the recording of polarity reversals in loess[20], we date each unit to within a precessional cycle rather than to the 1-thousand-year precision of the Chiloparts timescale.

The main palaeoanthropological limitation of the Loess Plateau is that it largely lacks evidence of the Early Palaeolithic, or associated hominin remains. The main exception is Lantian county in which a mandible and cranium of *H. erectus* have been found at Chenjiawo and Gongwangling, respectively. Both were dated by palaeomagnetism and correlation with the Luochuan loess–palaeosol sequence[21]. At Chenjiawo the mandible occurs in palaeosol S6 (684–710 thousand years ago), and at Gongwangling the layer containing the cranium is now correlated with S23 (1.59–1.65 Ma)[3]. Although some stone artefacts were found in Lantian county during the 1960s[22,23], they could not be assigned precise geological contexts. A few artefacts have recently been dated to the middle-to-late Pleistocene[24,25].

In our 2004–2017 investigations, we found a new Palaeolithic locality at Shangchen (34° 13′ 33″ N, 109° 28′ 39″ E) about 4 km north of Gongwangling. It is about 74 m thick and complete from L5 to L28 (equivalent to marine isotope stage 12 to marine isotope stage 80) (Fig. 2). After we found artefacts in loess and palaeosol sections, our main aim was to assign these a precise age. Particular attention was paid to evidence of multiple flaking, preferably from more than one direction; the presence of clear flake scars, percussion bulbs, retouch and secondary flake removals to distinguish artefacts from geofacts (Supplementary Information). In some sections, palaeomagnetic samples were taken next to where stone artefacts were discovered to confirm that they were in undisturbed deposits. Stone artefacts and mammalian fossils were also found at the base of S27 and the upper part of L28 in a small excavation in 2017.

Preliminary classifications of loess–palaeosol horizons are based on the identification of marker layers (palaeosol S5 and loess horizons L9, L15, L24 and L25) during our field investigations. Grain-size, chemistry and mineralogy analyses were performed to further distinguish sediment type (Extended Data Fig. 1). These indicate that the strata are aeolian, as are other typical Chinese loess deposits[7,8,10–16]. Rock magnetic measurements indicate that the sediments are suitable for palaeomagnetic dating (Supplementary Information and Extended Data Fig. 2). High-resolution magnetic susceptibility curves, generated from 1,076 samples with an approximately 6.9-cm average sampling interval, show variations that are the same as those of representative Chinese loess sections and confirm our pedostratigraphic identification. Systematic high-resolution palaeomagnetic measurements by progressive thermal demagnetization, based on 722 specimens with an approximately 10-cm average sampling interval, were used to determine the characteristic remanent magnetization direction, after removing one or two soft magnetization components (Extended Data Fig. 3).

[1]Key Laboratory of Ocean and Marginal Sea Geology, Guangzhou Institute of Geochemistry, Chinese Academy of Sciences, Guangzhou, China. [2]State Key Laboratory of Loess and Quaternary Geology, Institute of Earth Environment, Chinese Academy of Sciences, Xi'an, China. [3]Department of Archaeology, University of Exeter, Exeter, UK. [4]Key Laboratory of Vertebrate Evolution and Human Origins, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing, China. [5]Key Laboratory of Ocean and Marginal Sea Geology, South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou, China. [6]School of Geography and Environmental Engineering, Gannan Normal University, Ganzhou, China. [7]State Key Laboratory of Lithospheric Evolution, Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing, China. [8]College of Resources and Environmental Sciences, Hunan Normal University, Changsha, China. [9]Environmental Supervision Detachment of Nanning, Nanning, China. [10]Henan Institute of Geological Survey, Zhengzhou, China. [11]School of Geography, South China Normal University, Guangzhou, China. *e-mail: zhuzy@gig.ac.cn; r.w.dennell@exeter.ac.uk

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

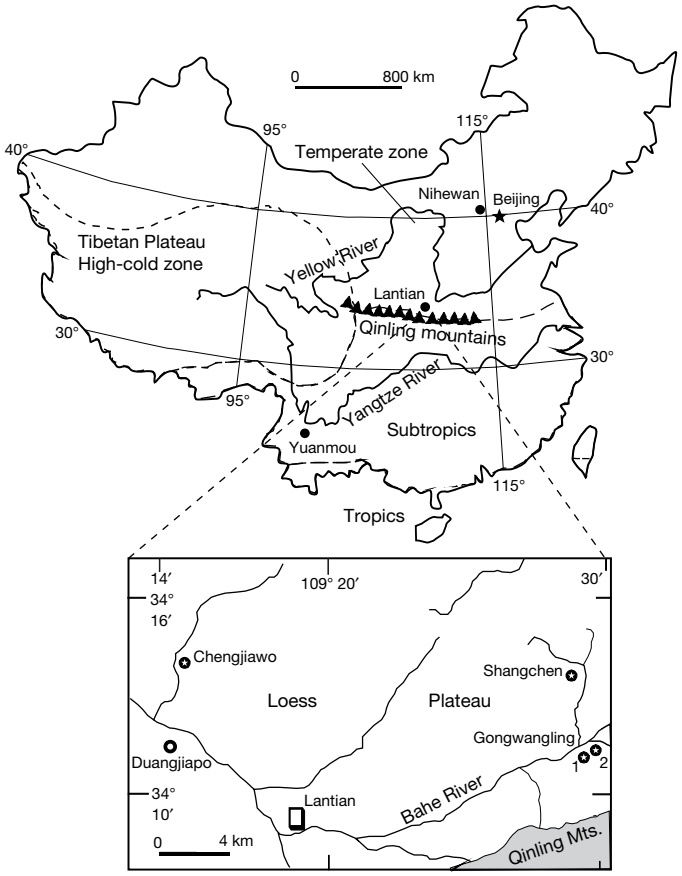


Fig. 1 | Location map of the Chinese Loess Plateau and the Lantian area. The Qinling Mountains form the climatic and biological boundary between the temperate north and subtropical south of China. The locations of major hominin sites in China, including the Nihewan basin and Yuanmou, are indicated. The Lantian basin lies in the southern Loess Plateau and north of the Qinling Mountains in the temperate zone of north China with Palaeolithic localities such as Gongwangling, Chenjiawo and Shangchen in the Lantian area. The main section of Shangchen is situated on denudated loess tablelands at an altitude of 898 m, along the north bank of the Bahe River. Duanjiapo is one of the typical sections of the Chinese loess–palaeosol sequence. Localities marked as 1 and 2 refer to Gongwangling 1 and Gongwangling 2, respectively.

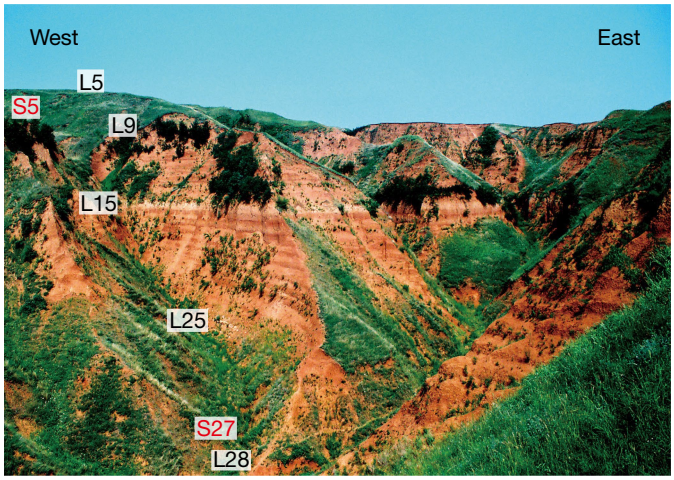


Fig. 2 | The landscape and loess–palaeosol section of the Shangchen Palaeolithic locality. The steep and dissected terrain exposes an aeolian loess–palaeosol sequence from L5 (412–479 thousand years ago) to L28 (2.11–2.13 Ma) that is fully developed, particularly with marker horizons such as the tripartite S5 with the reddest colour and highest magnetic susceptibility or the units L9 (869–943 thousand years ago), L15 (1.24–1.26 Ma) and L24–L25 (1.65–1.80 Ma), which represent particularly thick and silty loess (with coarse grain size) with a notable yellowish-white colour and lowest magnetic susceptibility. On the basis of these marker horizons, palaeomagnetic sampling and artefact collecting have been divided into five subsections along the continuously developed gully, and the main stratigraphic sequence has been established (see Supplementary Information and Extended Data Figs. 4, 5 for details). Many artefacts have been found in situ within the sampled area.

The magnetic polarity for each specimen was used to establish the local magnetostratigraphy (see Fig. 3, Supplementary Information, Extended Data Figs. 4, 5 and Source Data), and from this precise ages have been assigned to the artefacts in the loess–palaeosol sequence.

Eight major magnetozones are recorded in the Shangchen section (Fig. 3), four of which have normal polarity (N1 to N4) and four of which have reversed polarity (R1 to R4). By comparison with the geomagnetic polarity timescale[26,27], magnetozone N1 corresponds to the Brunhes chron, and the Matuyama/Brunhes boundary (0.78 Ma) is in horizons S7 to L8; magnetozone N2 corresponds to the Jaramillo subchron (0.99–1.07 Ma) and lies between horizons S10 and S12; magnetozone N3 corresponds to the Olduvai subchron (1.78–1.95 Ma), with its top and base lying within the lower part of L25 and the base of S26, respectively; and magnetozone N4 corresponds to the Réunion excursion (2.13–2.15 Ma) in L28. The four reversed polarity magnetozones are attributed to different periods of the Matuyama chron. A short-term polarity event, labelled e_1, might correspond to the Punaruu excursion at about 1.12 Ma[28] (see Supplementary Information). Stratigraphic continuity and the position of reversal boundaries in the Shangchen section indicate that the loess–palaeosol sequence is complete from L5 to L28 through a date range of approximately 0.41–2.16 Ma[10,27]. By using marker layers and palaeomagnetic reversal boundaries of offset sections[29] (see Supplementary Information and Extended Data Figs. 4, 5), the comprehensive main section and timescale of the Shangchen locality have been established from five subsections.

At Shangchen, 88 flaked and 20 unmodified stones were found in situ between S5 and L28. Here we discuss only those found in situ ($n = 96$) in layers S15 to L28, which date from 1.26 Ma to 2.12 Ma. The 82 flaked and 14 unmodified stones between S15 and L28 (Supplementary Table 6) were collected from 6 loess and 11 palaeosol strata within trenches cut for palaeomagnetic sampling, in situ at distances of less than 100 m from these trenches and exposed in the surfaces of 50–80° slopes, or within a 15.4-m^2 excavated area of S27–L28 (see Supplementary Information and Extended Data Figs. 6–10). Five categories of flaked stone were recognized: cores (17), flakes (9), flake fragments (8), fragments (14) and retouched pieces (34). The 34 retouched items were classified as scrapers (17), notches (4), scraper and/or notches (4), points and/or borers (6) and picks and/or bifaces (3). Cores were produced by bipolar ($n = 2$) and hard hammer percussion ($n = 15$), and those on elongate cobbles ($n = 9$) were most common. The average number of flakes per core was 4.6 (range of 1–7), from a mean of 2.3 directions (range of 1–6). As no natural stone exposures (for example, stream beds or outcrops) were encountered, isolated unmodified cobbles are regarded as manuports that may have been intended as cores or used as hammerstones. Two pieces identified as hammerstones show percussion damage. All artefacts except one were fresh, with no sign of rolling or abrasion. The most common raw materials (Supplementary Table 6) are different types of quartzite ($n = 72$) and quartz ($n = 20$) and their probable source is the foothills of the Qinling Mountains—which are situated 5–10 km to the south of the site—and the streams flowing from them. The horizon with the most stone items is S22 ($n = 33$, 1.54–1.57 Ma) (Supplementary Table 8 and Extended Data Fig. 6). A total of 28 stones (25 flaked and 3 unmodified) were found below S22 (Fig. 4). In S23 (1.59–1.65 Ma) a flake and a flake fragment were found; in S24 (1.71–1.73 Ma), a notch, scraper, core and two flake fragments; in L25 (1.73–1.80 Ma), a scraper, hammerstone, and two fragments; in S26 (1.89–1.95 Ma), a cobble; in L27 (1.95–2.09 Ma), a cobble and a fragment; and in S27 (2.09–2.12 Ma) (Extended Data Fig. 7), six artefacts were found (three cores, a core

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

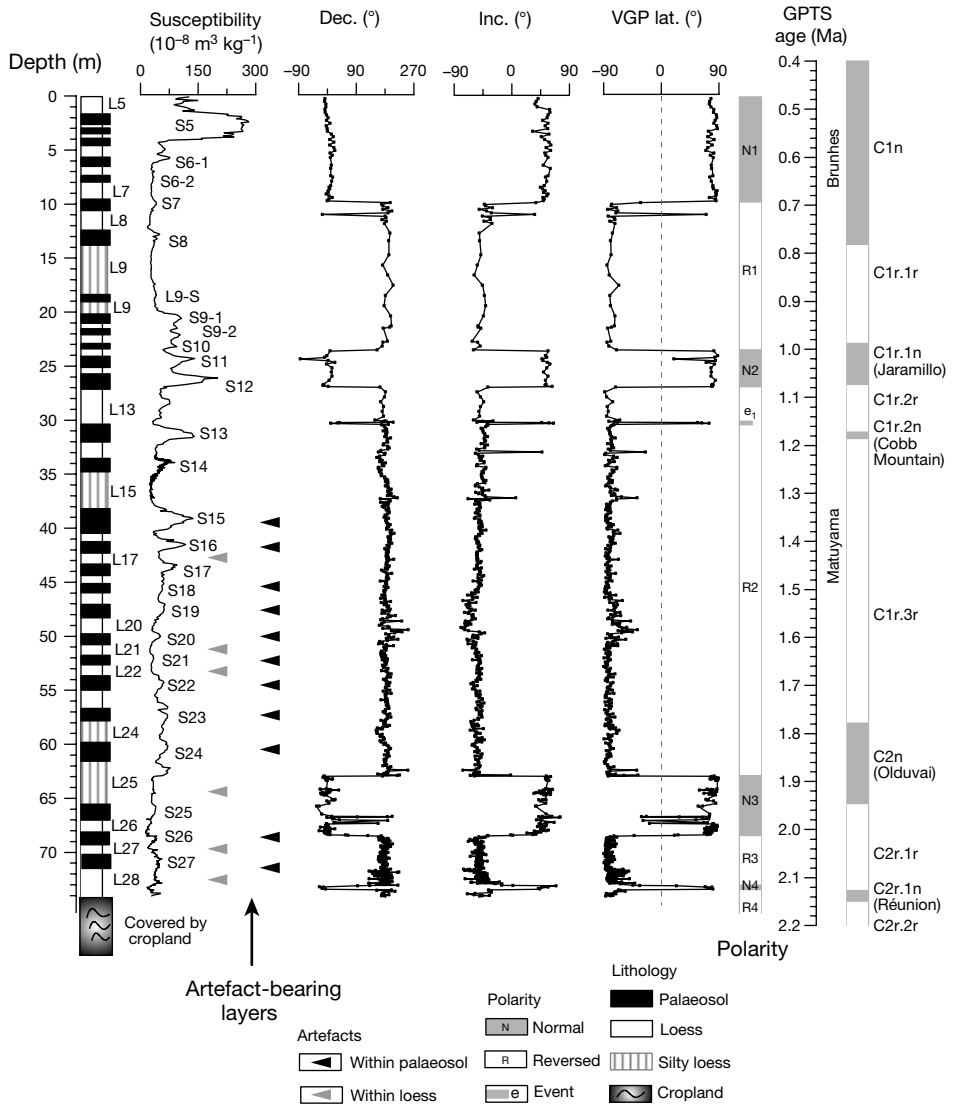


Fig. 3 | Summary of the loess–palaeosol section with artefact sequence of the Shangchen Palaeolithic locality. Magnetic susceptibility variations (high values in palaeosols and low values in loess), and the positions of key marker layers (S5, L9, L15 and L24–L25) are the same as in the standard Loess Plateau sections[7,8,10–16]. The geomagnetic polarity timescale is from previous publications[26,27]. Details of the palaeomagnetic data and artefacts within 17 excavation layers (S15, S16, S18–S24, S26 and S27, and L17, L21, L22, L25, L27 and L28) are provided in Supplementary Information, Extended Data Figs. 5–10 and Source Data. Artefacts found above S15 are not included in this paper because here we focus on the artefacts earlier than about 1.2 Ma. There are four normal polarity magnetozones, comprising N1 (chron C1n, 0–9.71 m), N2 (C1r.1n, 23.50–26.90 m), N3 (C2n, 62.94–68.44 m) and N4 (C2r.1n, 73.08–73.41 m). Four reversed polarity magnetozones, comprising R1 (9.71–23.50 m), R2 (26.90–62.94 m), R3 (68.44–73.08 m) and R4 (73.41–74.05 m), are attributed to different periods of the Matuyama chron. In addition, as an independent check on the Chiloparts timescale, we have estimated the ages of the main layers by using the average sedimentation rate of each segment to obtain the sedimentation rate age (see Supplementary Tables 2–5). The age difference between our sedimentation rate age and the Chiloparts timescale is small (~2.16%) and we therefore use the internationally recognized Chiloparts timescale throughout this paper. Dec., declination; GPTS, geomagnetic polarity timescale; Inc., inclination; VGP, virtual geomagnetic pole.

and/or scraper, a pointed piece and a flake tool). In the excavation into S27 and L28 (Extended Data Figs. 8, 9), a core and a flake tool were found at the S27/L28 boundary, and six artefacts were found in L28 (approximately 2.12 Ma): three scrapers, a point, a pointed piece and a flake. A mandibular fragment of a cervid, a bovid and other fossil bone fragments were also found (Extended Data Fig. 10). Using the Chiloparts timescale, the stone artefact and fossil horizon in the upper part of L28 between the Olduvai subchron and the Réunion excursion is dated to about 2.12 Ma. Details of artefact features are indicated in Supplementary Tables 6–8. Images of artefacts being extracted from original strata are shown in Extended Data Figs. 6–9, and laser three-dimensional scans of three pieces are shown in Supplementary Videos 1–3. All artefacts in these figures have clear flake features and evidence of repeated flaking from more than one direction. They are also consistent with the rest of the artefact sequence from Shangchen with respect to raw materials, flaking techniques and type of flaking product.

Four aspects of our discoveries at the Shangchen Palaeolithic locality are noteworthy. First, we have for the first time—to our knowledge—established the age of Early Pleistocene (below L15) artefacts in the Loess Plateau within a rigorous stratigraphic framework. Second, stone artefacts were found predominantly in 11 palaeosol layers (S15, S16, S18–S24, S26 and S27, with 80 items), which developed in a warm and wet climate, but in only 6 loess layers (L17, L21, L22, L25, L27 and L28, with 16 items), which indicate colder and drier conditions (Fig. 3). This pattern is consistent with that found in the loess–palaeosol sequence in Tajikistan[30]. Third, the 17 loess–palaeosol layers with artefacts span about 0.85 million years and indicate repeated—but not necessarily continuous—hominin occupation of the Chinese Loess Plateau between 1.26 and 2.12 Ma, equivalent to marine isotope stage

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Fig. 4 | Selected artefacts found in situ in layers S27–L28 (2.09–2.12 Ma), L27 (1.95–2.09 Ma), L25 (1.73–1.80 Ma) and S23 (1.59–1.65 Ma) from the Shangchen Palaeolithic locality. a, SC 20120507-3, a flake tool from S27. **b**, SC 20120507-2, a pointed piece from S27. **c**, SC 20120516-1, a core from S27. **d**, SC 20120507-1, a core from S27. **e**, SC-B D2-2, a bipolar fragment from L27. **f**, SC 20120502-6, a flake from S23. **g**, 2010-06 SC-L25, a scraper from L25. **h**, SC-K5, a flake tool from S27–L28. **i**, SC-K4, a core from S27–L28. **j**, SC070926-1, a flake fragment from S23. Artefacts from S23 are similar in age to the the Gongwangling *H. erectus* cranium[3]. Each gradation in the scales represents 1 cm. Artefacts SC-K4 and SC-K5 are also shown in Extended Data Fig. 10. SC, abbreviation for Shangchen site.

39 to marine isotope stage 80. Fourth, and most importantly, the oldest artefact age of approximately 2.12 Ma at Shangchen implies that hominins had left Africa before the date suggested by the earliest evidence from Dmanisi (about 1.85 Ma). This makes it necessary to reconsider the timing of initial dispersal of early hominins in the Old World.

Data availability

The authors declare that all data supporting the findings of this study are available within the paper and its Supplementary Information. Source Data for Fig. 3 are provided with the paper.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0299-4.

Received: 9 December 2016; Accepted: 30 May 2018;
Published online 11 July 2018.

1. Ferring, R. et al. Earliest human occupations at Dmanisi (Georgian Caucasus) dated to 1.85–1.78 Ma. *Proc. Natl Acad. Sci. USA* **108**, 10432–10436 (2011).
2. Zhu, R. X. et al. Early evidence of the genus *Homo* in East Asia. *J. Hum. Evol.* **55**, 1075–1085 (2008).
3. Zhu, Z. Y. et al. New dating of the *Homo erectus* cranium from Lantian (Gongwangling), China. *J. Hum. Evol.* **78**, 144–157 (2015).
4. Zaim, Y. et al. New 1.5 million-year-old *Homo erectus* maxilla from Sangiran (Central Java, Indonesia). *J. Hum. Evol.* **61**, 363–376 (2011).
5. Zhu, R. X. et al. New evidence on the earliest human presence at high northern latitudes in northeast Asia. *Nature* **431**, 559–562 (2004).
6. Ao, H., Dekkers, M. J., Wei, Q., Qiang, X. & Xiao, G. New evidence for early presence of hominids in North China. *Sci. Rep.* **3**, 2403 (2013).
7. Liu, T. S. et al. *Loess and the Environment* (China Ocean, Beijing, 1985).
8. Rutter, N., Ding, Z., Evans, M. E. & Liu, T. Baoji-type pedostratigraphic section, Loess Plateau, north-central China. *Quat. Sci. Rev.* **10**, 1–22 (1991).
9. Lisiecki, L. E. & Raymo, M. E. A Pliocene-Pleistocene stack of 57 globally distributed benthic $\delta^{18}O$ records. *Paleoceanography* **20**, PA1003 (2005).
10. Ding, Z. L. et al. Stacked 2.6-Ma grain size record from the Chinese loess based on five sections and correlation with the deep-sea ^{18}O record. *Paleoceanography* **17**, 5-1–5-21 (2002).
11. Heller, F. & Liu, T.-S. Magnetostratigraphical dating of loess deposits in China. *Nature* **300**, 431–433 (1982).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

12. Kukla, G. & An, Z. Loess stratigraphy in central China. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **72**, 203–225 (1989).
13. Zheng, H., An, Z. & Shaw, J. New contributions to Chinese Plio-Pleistocene magnetostratigraphy. *Phys. Earth Planet. Inter.* **70**, 146–153 (1992).
14. Heslop, D., Langereis, C. G. & Dekkers, M. J. A new astronomical timescale for the loess deposits of Northern China. *Earth Planet. Sci. Lett.* **184**, 125–139 (2000).
15. Guo, Z. T., Jiang, W. Y., Lu, H. Y., Wu, N. Q. & Yao, X. F. Pleistocene climate extremes in East Asia and their causes. *Earth Sci. Frontiers* **9**, 113–120 (2002).
16. Liu, Q., Roberts, A. P., Rohling, E. J., Zhu, R. & Sun, Y. Post-depositional remanent magnetization lock-in and the location of the Matuyama–Brunhes geomagnetic reversal boundary in marine and Chinese loess sequences. *Earth Planet. Sci. Lett.* **275**, 102–110 (2008).
17. Gibbard, P. & Cohen, K. M. Global chronostratigraphical correlation table for the last 2.7 million years. *Episodes* **31**, 243–247 (2008).
18. Cohen, K. M. & Gibbard, P. L. Global chronostratigraphical correlation table for the last 2.7 million years. http://www.stratigraphy.org/index.php/ics-chart-timescale (2016).
19. Liu, Q. et al. Magnetostratigraphy of Chinese loess–paleosol sequences. *Earth Sci. Rev.* **150**, 139–167 (2015).
20. Sun, Y., Clemens, S. C., An, Z. & Yu, Z. Astronomical timescale and palaeoclimatic implication of stacked 3.6-Myr monsoon records from the Chinese Loess Plateau. *Quat. Sci. Rev.* **25**, 33–48 (2006).
21. An, Z. & Ho, C. New magnetostratigraphic dates of Lantian *Homo erectus*. *Quat. Res.* **32**, 213–221 (1989).
22. Dai, E.-J. & Chi, H.-G. Discovery of palaeoliths at Lantian, Shensi. *Vertebrata PalAsiatica* **8**, 152–156, 161 (1964).
23. Tai, E. C. & Hsu, C. H. New finds of palaeoliths from Lantian. *Acta Archaeologica Sinica* **1**, 1–12 (1973).
24. Wang, S. et al. Newly discovered Palaeolithic artefacts from loess deposits and their ages in Lantian, central China. *Chin. Sci. Bull.* **59**, 651–661 (2014).
25. Zhuo, H. X. et al. Chronology of newly-discovered paleolithic artifact assemblages in Lantian (Shaanxi province), central China. *Quat. Res.* **86**, 316–325 (2016).
26. Cande, S. C. & Kent, D. V. Revised calibration of the geomagnetic polarity timescale for the Late Cretaceous and Cenozoic. *J. Geophys. Res.* **100**, 6093–6095 (1995).
27. Ogg, J. G. in *The Geologic Time Scale 2012* (eds Gradstein, F. M. et al.) 85–114 (Elsevier, Oxford, 2012).
28. Roberts, A. P. Geomagnetic excursions: knowns and unknowns. *Geophys. Res. Lett.* **35**, L17307 (2008).
29. Kappelman, J. et al. in *Geology and Paleontology of the Miocene Sinap Formation, Turkey* (eds Fortelius, M. et al.) 41–66 (Columbia Univ. Press, New York, 2003).
30. Ranov, V. The 'loessic palaeolithic' in South Tadjikistan, central Asia: its industries, chronology and correlation. *Quat. Sci. Rev.* **14**, 731–745 (1995).

Acknowledgements This research was supported by the National Basic Research Program of China (Grant 2010CB833400), projects of NSFC (grants 41102115 and 41662012) and Projects of Chinese Academy of Sciences (grants KZCX2-SW-133, KZCX3-SW-152, 2013TIZ0008, XDB26000000, SKLLQG1525, SKLLQG1502, SKLLQG1501 and SKLLQG1122). This is contribution number IS-2546 from GIGCAS and Key Deployment Projects of IVPPCAS. We thank Z. An, R. Zhu, Z. Ding, Z. Guo, Z. Qiu, W. Liu, Y. Pan, W. Dong, H. Tong, H. Zheng, X. Tan, X. Qiang, H. Lu, Y. Pan and C. Deng for guidance in loess–palaeosol stratigraphy, palaeomagnetism and palaeoanthropology; the government of Lantian County and Cultural Relic Management Department of Gongwangling for helping our fieldwork, as well as our colleagues, Y. Kuang, Y. Han, S. Qin, H. Huang, S. Peng, M. Li, Z. Ruan, R. Deng, Y. Hao, Y. Chen, W. Chen, F. Li and Z. Li; and L. Hurcombe for advice on the artefacts.

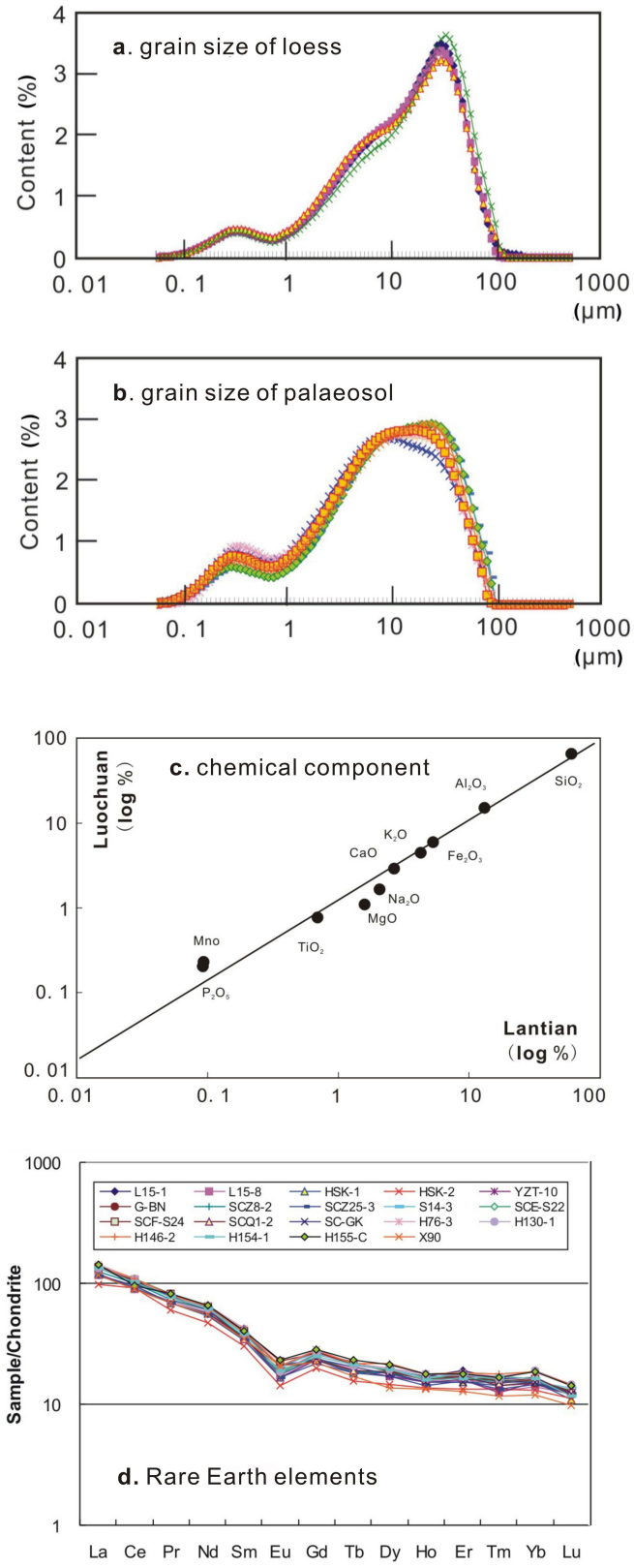
Reviewer information *Nature* thanks J. Kappelman, M. Petraglia, A. Roberts and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Z.Z. directed the research plan, participated in all investigations from 2004 to 2017, and co-authored the text, the Supplementary Information, and the Extended Data legends. W.H. is the principal archaeologist, discoverer of the Gongwangling cranium and artefacts in 1964, and also authenticated the artefacts reported in this paper. R.D. participated in the fieldwork, found many artefacts in the field, confirmed all field observations, authenticated the artefacts reported in this paper, and co-authored and edited the text and Supplementary Information. Y.W., S.Q., Z.R., J.X. and J.H. contributed to the major field investigations and found several artefacts, assisted with laboratory measurements, and writing and editing of this paper. S.Y. and Y.H. participated in part of the fieldwork and assisted in authenticating and measuring all artefacts reported in this paper. T.O. participated in the palaeomagnetic measurements.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0299-4.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0299-4.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to Z.Z. or R.D.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

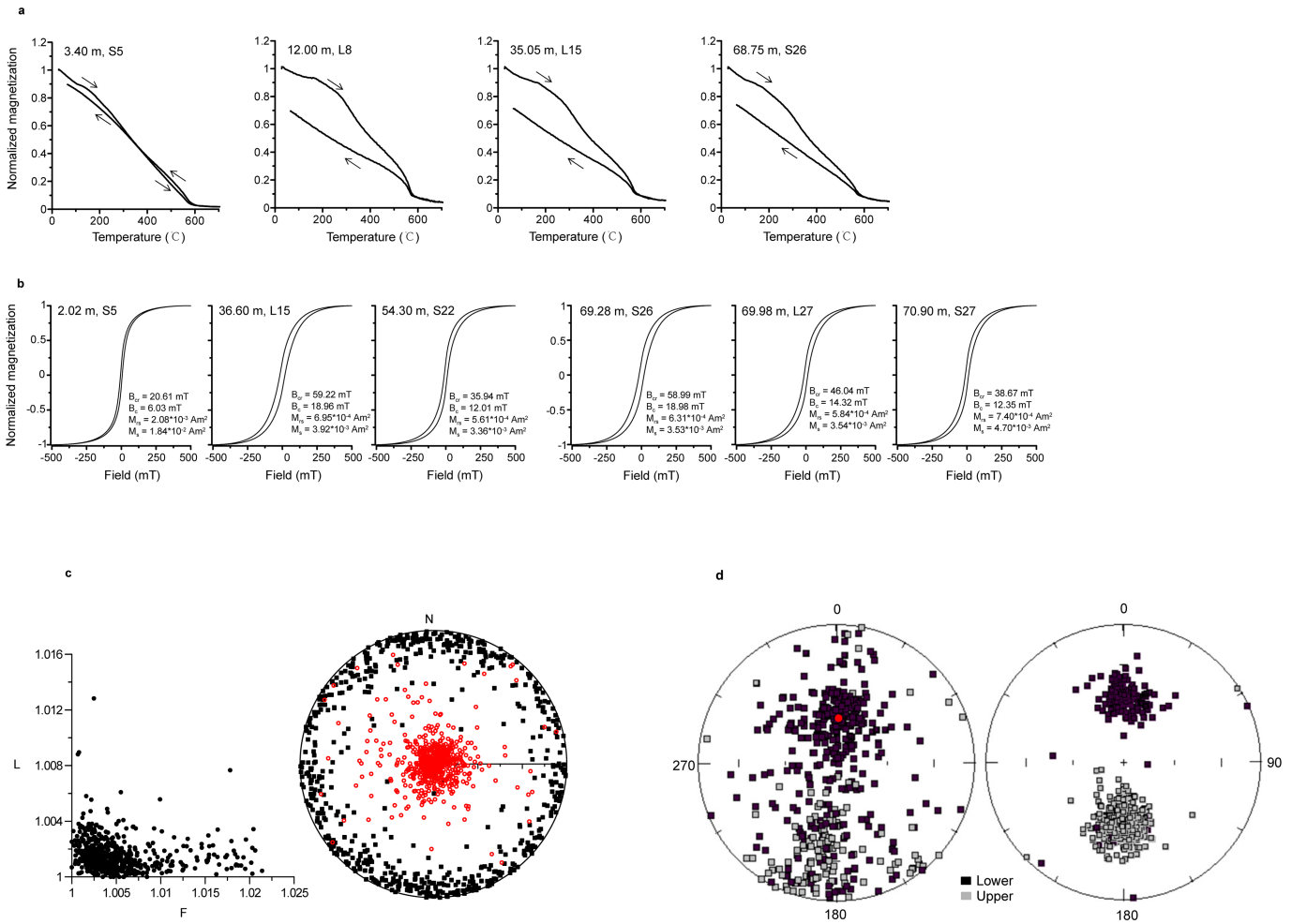
© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 1 | Material composition of aeolian deposits in the Lantian area. a, **b**, Grain size distributions for loess (**a**) and palaeosol (**b**) samples from the Shangchen section. The features of sediments in Shangchen section are the same as those from the Lingtai section[31], which is one of the typical Chinese loess sections. This indicates that sediments from the Shangchen section are aeolian deposits that are similar to those at Lingtai. **c**, Comparison of major elements between the loess sediments at Lantian and Luochuan[7]. The *x* and *y* coordinates represent logarithmic values of the major element content of loess from Luochuan and Lantian, respectively, and the plot shows that loess from the two sites is geochemically similar (data fall on or close to the 1:1 line). **d**, Chondrite-normalized rare-earth element distribution patterns for loess and palaeosol samples from the Shangchen section. Characteristics of partition modes of rare-earth elements from loess and palaeosol samples in the Shangchen section indicate they are the same as those from the Luochuan section.

31. Sun, Y. B., Lu, H. Y. & An, Z. S. Grain size distribution of quartz isolated from Chinese loess/paleosol. *Chin. Sci. Bull.* **45**, 2296–2298 (2000).

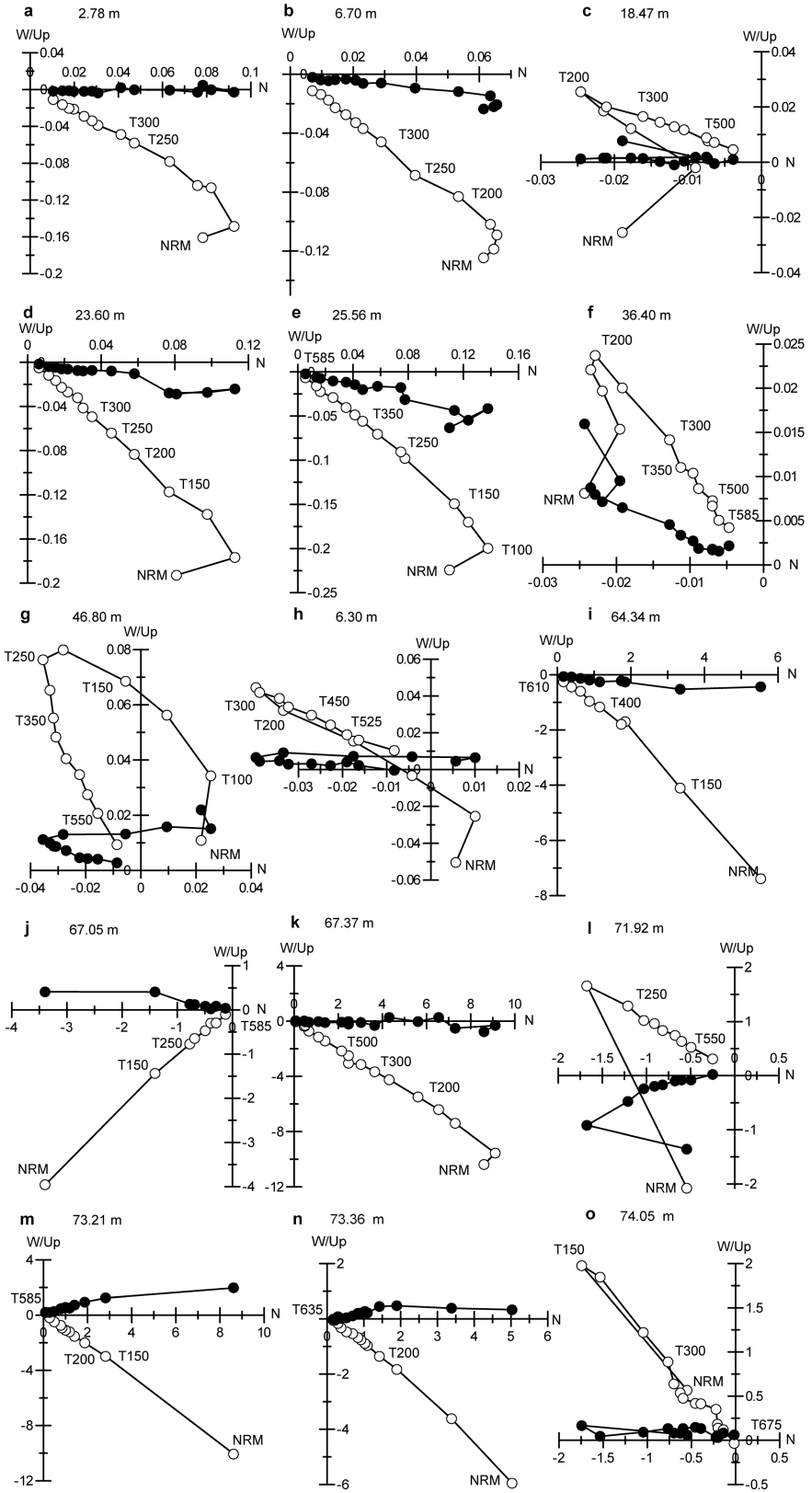
© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 2 | Rock magnetism. a, Temperature-dependent magnetization variations for four representative samples from the Shangchen section. Arrows indicate heating or cooling runs. These curves were obtained in air using a field of 100 mT, and indicate the presence of maghemite, dominant magnetite and haematite. This type of magnetic mineral assemblage is typical of that previously recovered for the Chinese loess–palaeosol sequence[19,32]. **b**, Hysteresis loops for representative samples after correction for paramagnetic slope. M_s, saturation magnetization; M_{rs}, saturation remanence; B_c, coercive force; and B_{cr}, coercivity of remanence. **c**, Anisotropy of magnetic susceptibility for 694 specimens from loess L5 to loess L28 in the main section that we studied, and parallel sections at the Shangchen locality. Left, Flinn diagram. L, lineation ($\kappa_{max}/\kappa_{int}$). F, foliation ($\kappa_{int}/\kappa_{min}$). Right, stereographic projections of anisotropy-of-magnetic-susceptibility ellipsoids of specimens. Data for κ_{max} and κ_{min} are shown as black squares and red circles, respectively. The data are indicative of normal, undisturbed sedimentary fabrics and support the magnetostratigraphic interpretation presented in this study. **d**, Equal area projections for the natural remanent (left) and characteristic remanent (right) magnetization directions for all 694 studied specimens.

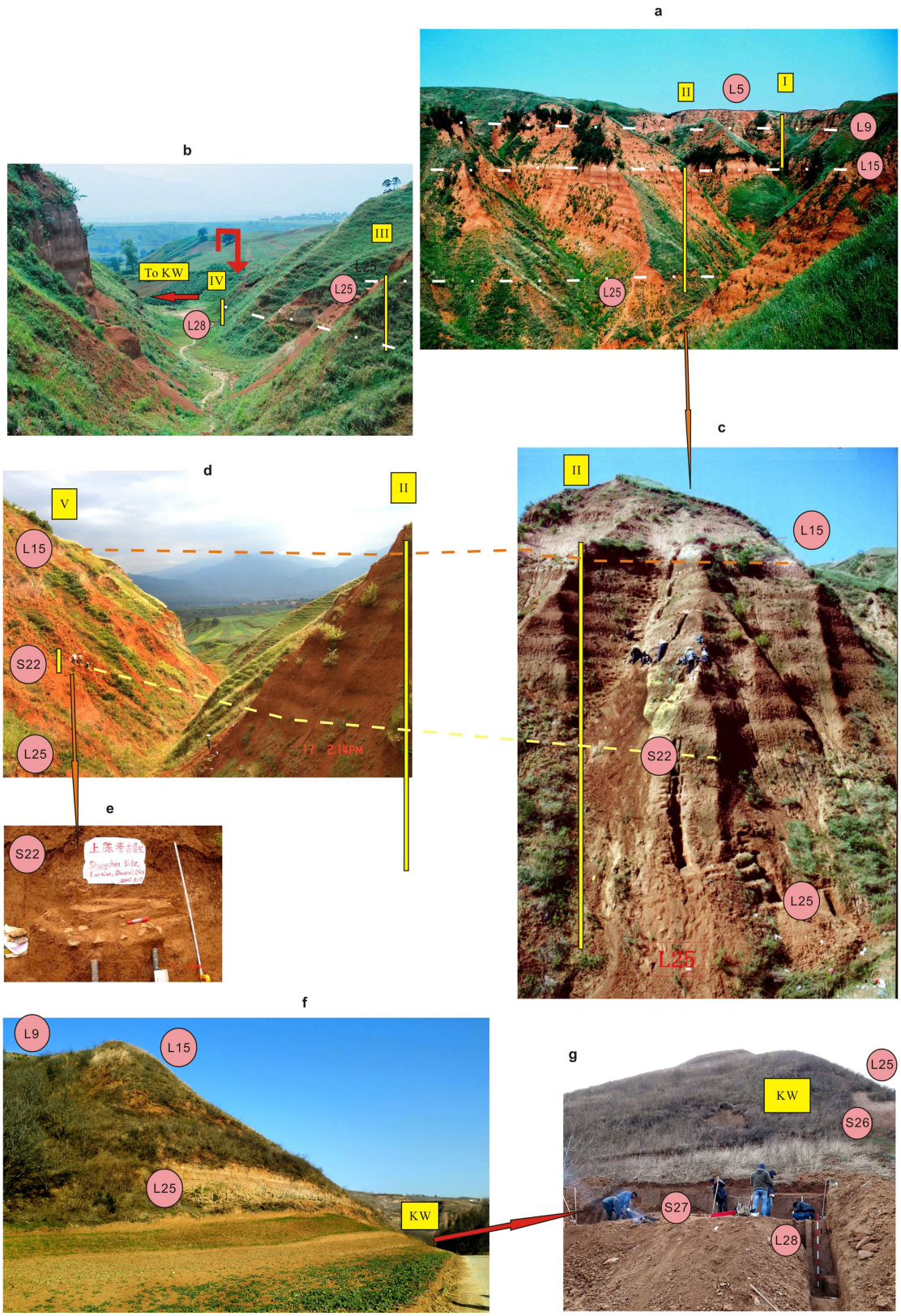
32. Liu, Q., Deng, C., Torrent, J. & Zhu, R. Review of recent developments in mineral magnetism of Chinese loess. *Quat. Sci. Rev.* **26**, 368–385 (2007).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 3 | Palaeomagnetic dating. a–o, Demagnetization diagrams for representative specimens from the Shangchen section. Solid and open symbols refer to data projected onto the horizontal and vertical planes, respectively. Scales of demagnetization temperature and coordinate axes are in degrees Celsius and mA m^{-1}, respectively. In **m**, **n**, the specimens from the magnetic horizon of the Réunion excursion are shown. See Source Data for detailed listings of palaeomagnetic data from the Shangchen section.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

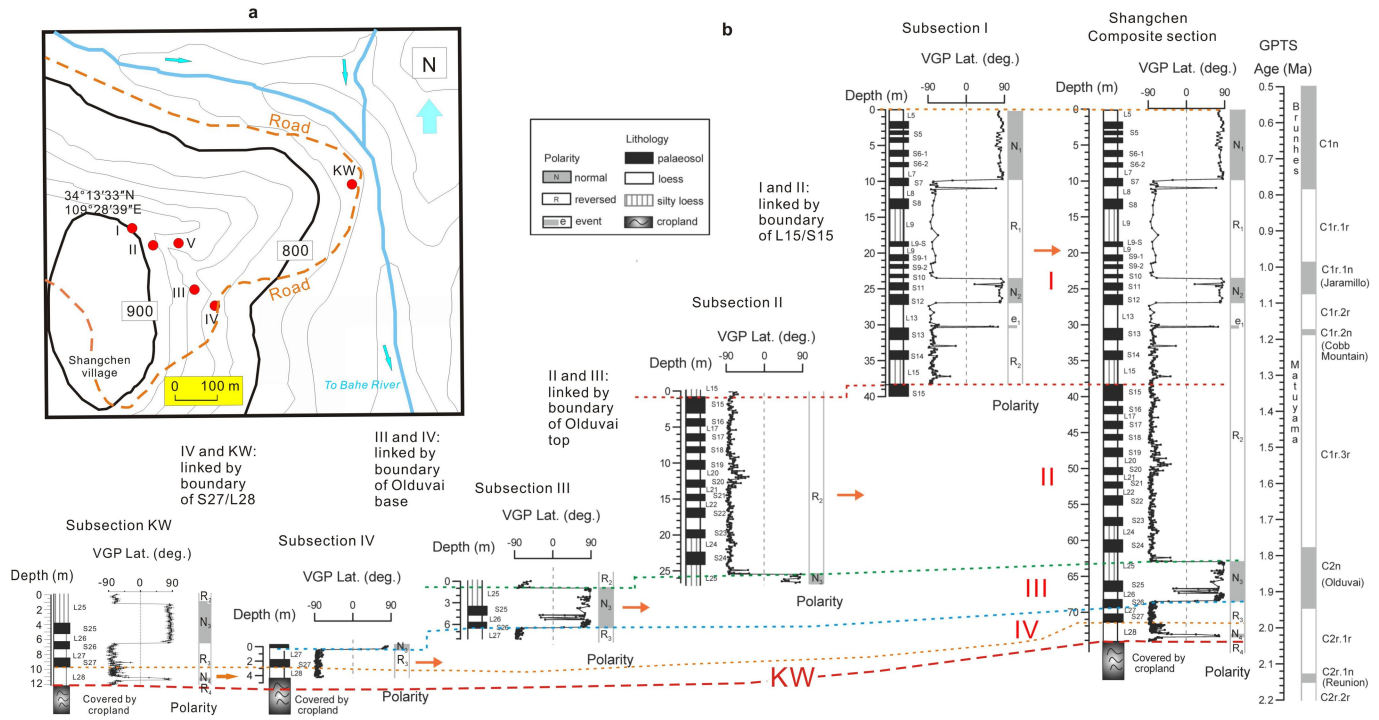


Extended Data Fig. 4 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

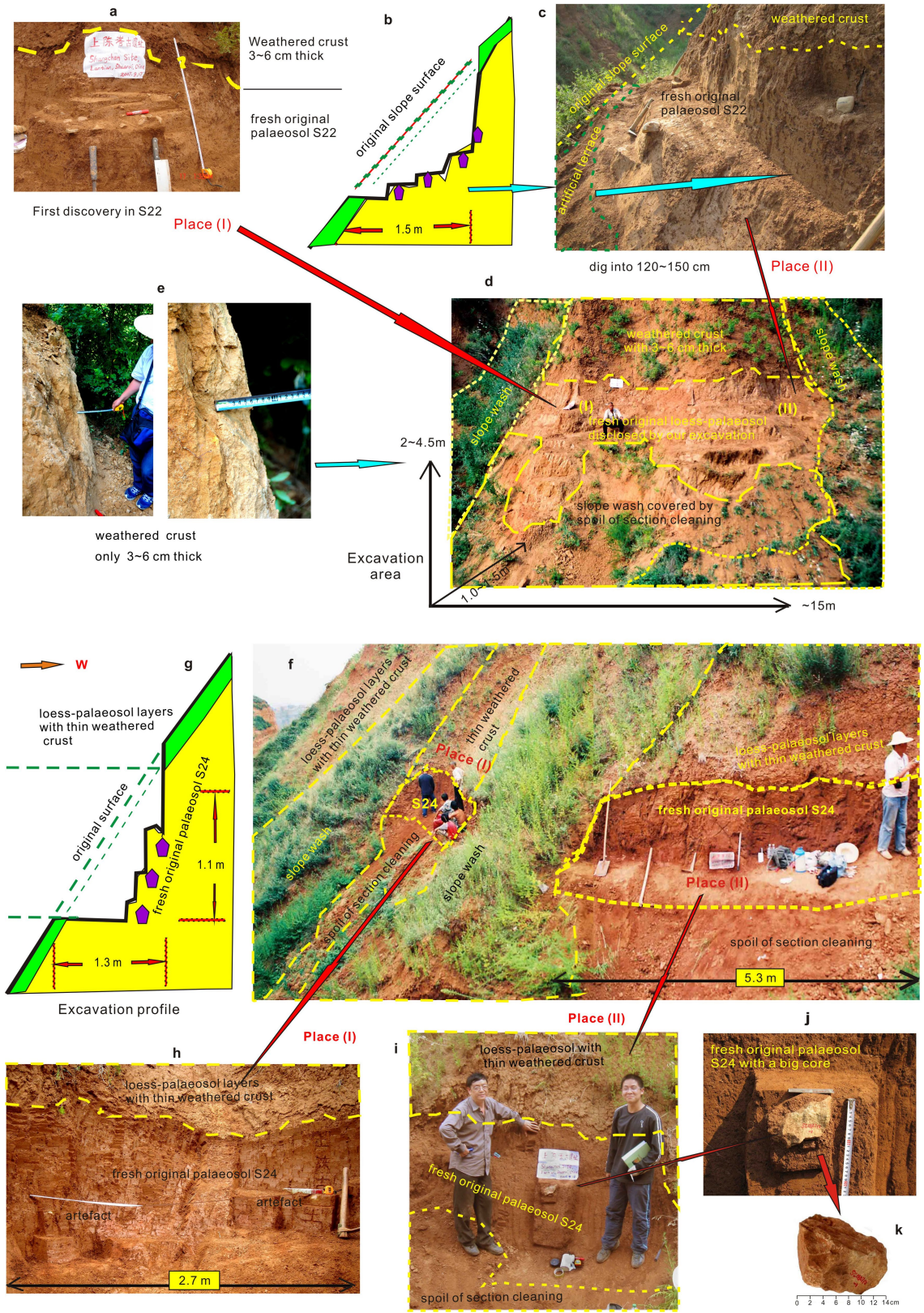
Extended Data Fig. 4 | Landscape in which palaeomagnetic sampling and artefact collecting were carried out at the Shangchen locality. The establishment of magnetostratigraphy in the Shangchen locality was based on the traditional and effective-linking methods (using marker layers and palaeomagnetic reversal boundaries of offset sections[29]). The comprehensive main section and timescale of the Shangchen locality have been established from five subsections, the spatial distribution of which is shown below. **a**, **b**, The main sections are exposed continuously along the same gully. Because of the steep terrain and multistep gully bottom (see **c**, **d**), we divided the sampling into four subsections (offset sections) in this gully. Several marker layers—such as L9, L15 and L25—that are white and thick are obvious in the section at short range and can easily be used to link the subsections. The subsections I, II, III and IV contain the following layers: I (L5–L15), II (L15–L25), III (L25–L27), and IV (L27-L28, on the other side of the slope). **c**, The steep subsection II with fresh outcrops of original loess and palaeosols. Many artefacts were found within more than 10 horizons in subsection II or on both its sides. The first artefact was found here in S22. Note the two individuals on the upper part of the section as an approximate scale. **d**, Subsection V, which is a short parallel section opposite subsection II (**c**). **e**, Many artefacts were found in situ in palaeosol S22 during our sampling and excavation. Palaeomagnetic analysis confirms that the artefacts were collected from undisturbed loess or palaeosol (see Supplementary Information). The steepness of slope (shown in **c**, **d**) prevents large-scale excavation. More details can be seen in Extended Data Fig. 6. **f**, **g**, The section of the exploratory trench (named subsection KW) contains layers L25–L28. It is part of the same small hill as the main composite section (subsections I–IV), and is located about 500 m northeast of palaeomagnetic section IV (see Extended Data Fig. 5a). Based on the outcrops of marker layers L9, L15 and L25 in the loess–palaeosol sequence, these sections can easily be linked (see Extended Data Fig. 5 for further details).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 5 | Distribution and linking method of sections at the Shangchen locality. a, Schematic plan of the subsections of palaeomagnetic sampling. The numbers 800 and 900 refer to the height in metres above sea level. The contour interval is 20 m. Subsection KW refers to the additional parallel section from L25 to L28 with the excavation into S27 and L28. **b**, Schematic of the method used to link the five subsections in the Shangchen locality. This figure and Extended Data Fig. 4 show the loess–palaeosol stratigraphy, with key marker layers, sampled subsections and the locations where artefacts were found. The details of the method used to link the five subsections are given below. Subsections I and II are linked on the basis of the L15/S15 boundary. Measured horizons of subsection I start from the mid-to-lower part of L5 and run downward to the base of marker layer L15 (to the boundary of L15/S15). Measured horizons of subsection II start from the lowest part of L15 and run downward to the lower part of L25. Because the measured thickness of marker layer L15 is the same in both subsections I and II, we cut out the overlapped L15 in subsection II. Thus, subsection I ends at the base of L15, and subsection II begins at the top of S15. Subsections II and III are linked on the basis of the top boundary of the Olduvai subchron. A small normal-polarity segment of the Olduvai subchron was detected in the marker layer L25 at the base of II. Measured horizons of subsection III start from the lower part of L25 to the top of L27, and record the entire Olduvai subchron with a small reversed-polarity segment on the top (in L25) and at the base (top of L27) of the subsection. Thus, the overlapped normal-polarity segment of the Olduvai subchron at the base of subsection II and a small reverse-polarity segment on top of subsection III were cut out. The reverse-polarity segment at the base of subsection II is linked to the normal-polarity top segment of subsection III, and the whole normal-polarity segment of the Olduvai subchron is contained within subsection III. Subsections III and IV were linked on the basis of the base of the Olduvai subchron. Measured horizons of subsection IV start from the top of L27 to the mid-to-upper part of L28, with a small normal-polarity segment of the base of the Olduvai subchron at the top. After cutting out both the small reverse-polarity segment at the base of subsection III and the small normal-polarity segment on top of subsection IV, the reverse-polarity segment just below the Olduvai subchron in subsection IV is linked to the normal-polarity base of the Olduvai segment in subsection III. Subsection KW contains horizons of L25 (marker layer) to L28, including units L25, S25, L26, S26, L27, S27 and L28. Against the general background of reversed magnetostratigraphy, two normal magnetozones—the Olduvai subchron (1.78–1.95 Ma) and the Réunion excursion (2.13–2.15 Ma)—are recorded. Therefore, section IV and section KW are linked by the base of S27 (that is, S27 at the base of section IV is linked to L28 at the top of subsection KW). Subsection V is opposite of subsection II, which is on the other side of a narrow gully and has two distinct marker layers, L15 in the upper half and L25 in the lower half. There is a palaeosol layer with many artefacts in the middle part of subsection V. The stratigraphic correlation and horizon number demonstrate that this layer belongs to S22. Palaeomagnetic measurements on samples collected from this parallel section all showed reversed polarity.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

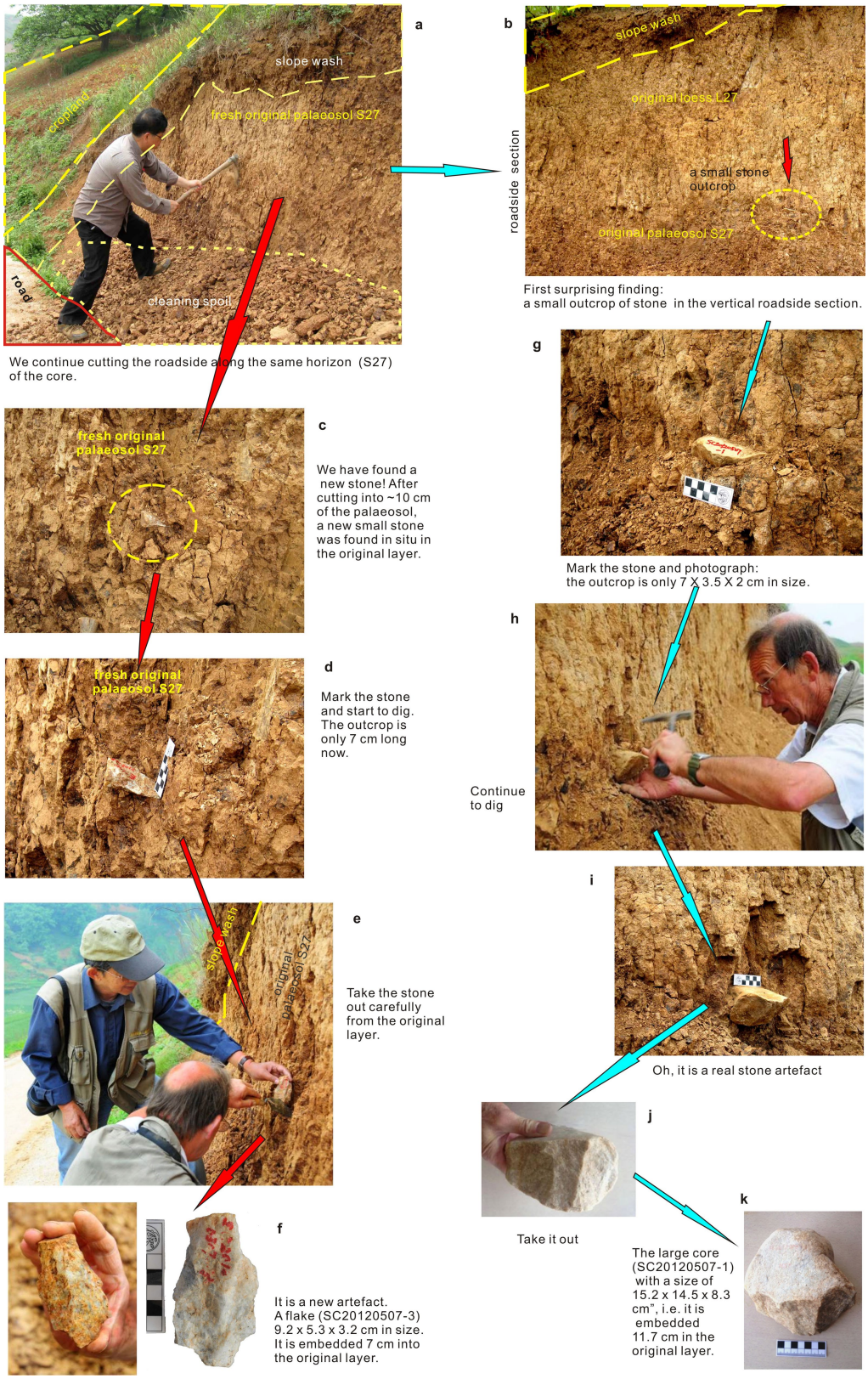


Extended Data Fig. 6 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

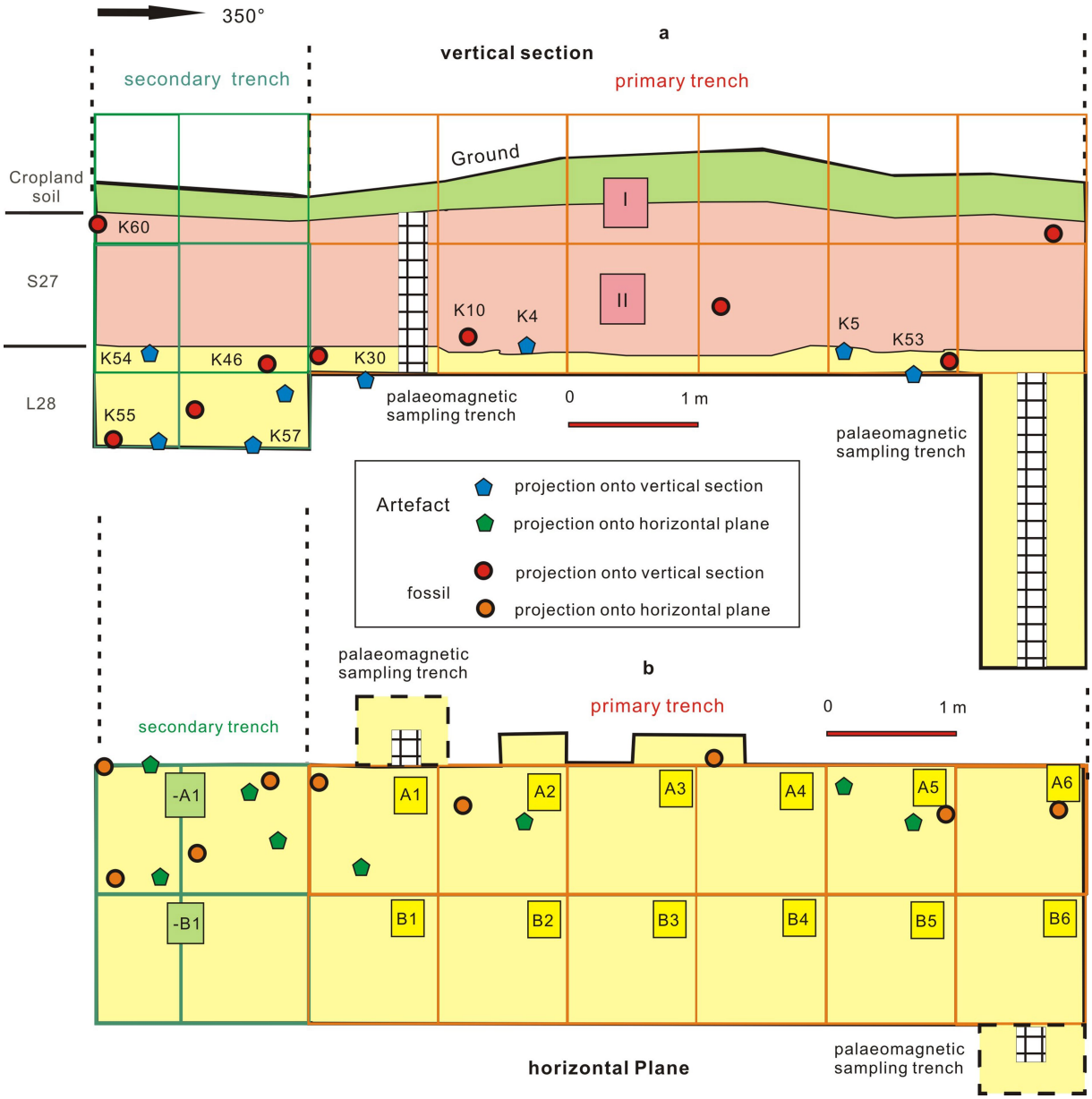
Extended Data Fig. 6 | Stone artefacts found during the sampling of S22 (1.54–1.57 Ma) and S24 (1.71–1.73 Ma) at the Shangchen section. **a**, First discovery of a stone artefact in S22 ('place I'). **b**, Schematic profile of the section-cutting of place II. **c**, The place II section was excavated into the original palaeosol layer to a depth of about 150 cm. Note the presence of stone artefacts on a ledge in the original palaeosol, below the weathered crust. **d**, View of the excavation in subsection V. **e**, Profile of the original slope surface with a weathered crust of only 3–6-cm thickness, and no slope wash. Palaeomagnetic analysis of the samples taken around the stone artefacts confirmed that they were in undisturbed deposit. **f**, View of the excavation in S24. Note the weathered surface crust upslope, the slope wash to the left of the cutting and section cleaning debris below the excavation. **g**, Schematic profile of excavation into S24. **h**, Side view of place I of the section cut by the excavation, showing two artefacts within fresh original palaeosol. Note that there is a clear difference between the thin weathered crust and the original loess. **i**–**k**, Discovery and exposure of a large quartzite core in place II of the section (**i**, **j**), which led to the removal of the core shown in **k**. Details of the large core (SC080710-1) are given in Supplementary Table 6. Analysis of the palaeomagnetic samples taken alongside the stone artefacts showed that the loess and palaeosol was undisturbed, and did not represent slope wash.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



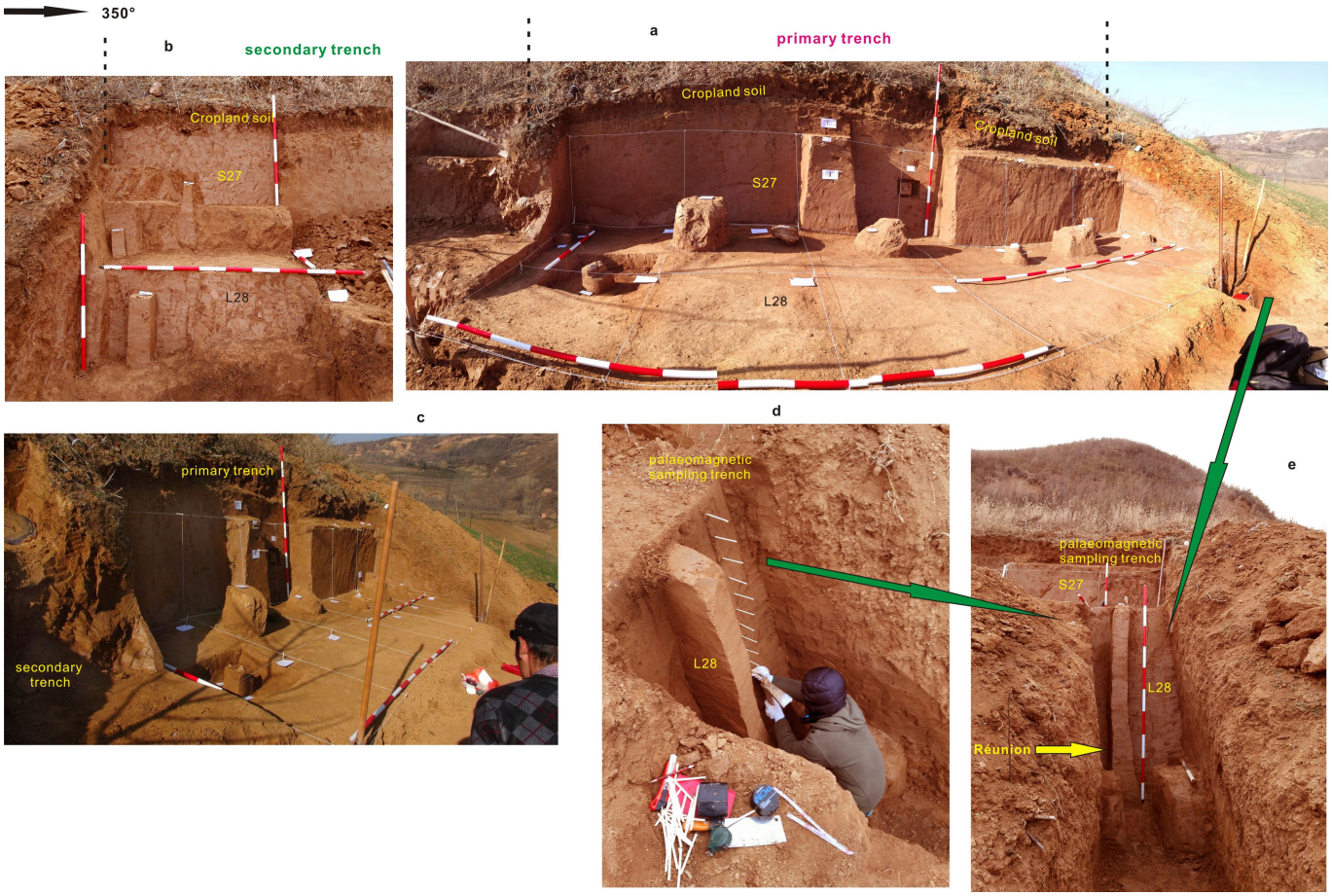
Extended Data Fig. 7 | Details of artefacts excavated from the roadside section at the Shangchen locality from S27, which occurs below the Olduvai subchron and is dated to approximately 2.1 Ma. The hill slope is covered by a dark brown slope wash of 10–30-cm thickness that is easily distinguishable from the in situ loess in the roadside section. After cleaning the section, stone artefacts were exposed between 1.5 and 3 m below the top of the section, and extracted from in situ loess. **a**, **b**, This shows the obvious difference between slope wash and the original loess–palaeosol layer. **b**, First finding of a stone standing out of the surface crust which is situated about 1.5 m below the top of slope wash. **c**–**f**, The finding and collecting process for the artefact is the same in Extended Data Fig. 6. Details of the artefact SC 20120507-3—which was found after digging >30 cm into the fresh original palaeosol layer—are given in Fig. 4a, Supplementary Tables 6, 7 and Supplementary Video 1. **g**–**i**, A stone is marked with its specimen number before removal (**g**), and the artefact is disclosed step by step (**h**, **i**). **j**, The stone artefact after removal. **k**, View of core (SC20120507-1). Clear flake scars are evident. This piece is shown in Fig. 4d and details are given in Supplementary Tables 6, 7.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



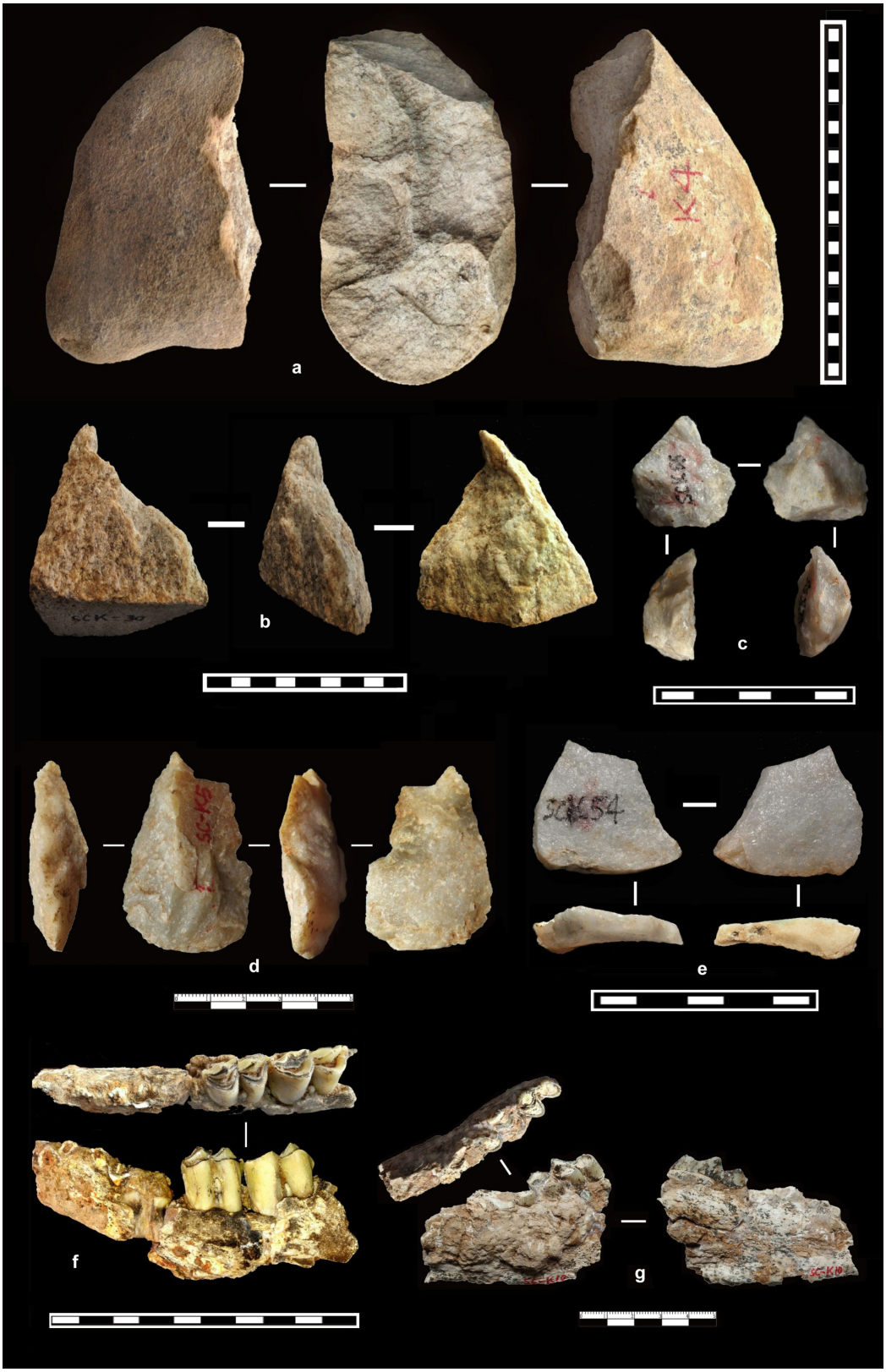
Extended Data Fig. 8 | The stratigraphic partition, grid layout and distribution of artefacts and fossils in the exploratory trench (subsection KW) in S27 and L28. a, **b**, The vertical section (**a**) and the horizontal plane (**b**) of the exploratory trench, which is composed of a primary and a secondary trench. The primary exploratory trench is 6 m long, 1.4–1.78 m deep and 2 m wide, and the secondary trench is 1.7 m long, 2 m deep and 2 m wide. The excavated area (approximately 15.4 m^2) was divided into a grid of 1-m × 1-m squares. I, II, the numbers of grid lines on the vertical section. A1–A6 and B1–B6, numbers of the grid squares on the horizontal plane in the primary trench. –A1, –B1, the numbers of grid lines on the horizontal plane in the secondary trench. In the south and the north of the primary trench, 108 palaeomagnetic samples were continuously collected in two sampling trenches. The north sampling trench is 0.8–1 m wide and 2.3 m deep, and the south sampling trench is 0.4 m wide and 1.2 m deep. Some stone artefacts and fossil remains were excavated in situ in S27 and L28 from our exploratory trench. The features of eight artefacts can be seen in Fig. 4, Extended Data Fig. 10 and Supplementary Table 6. The fossils from the excavation were primarily tooth and bone fragments (Extended Data Fig. 10) of different animals, including a cervid, a bovid and a suid. No bones or jaws were complete. The fossil remains and stone artefacts occurred on the same horizon and were close to one another. Note that the artefacts and fossils also occur in the deepest part of the trench and not near the modern slope surface.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 9 | The excavation of the exploratory trench (the subsection KW) in layers S27 and L28 at Shangchen locality. **a**–**c**, The primary (**a**, **c**) and secondary (**b**) exploratory trenches. **d**, **e**, Palaeomagnetic sampling trench in the north of the primary exploratory trench. The boundary between the cultivated horizon (cropland soil) and original loess horizon (S27) is very clear. The top layer is a grass-covered cropland soil of a depth of 20–40 cm (without any crops planted at this time). Underlying the cultivated layer are clear, homogeneous, solid and undisturbed loess–palaeosol strata. Several stones and fossils were found in situ in the original palaeosol–loess strata (S27–L28). These stones and mammalian fossils were kept in the reserved pillars of palaeosol and loess during the exploration process (**a**–**c**). The Réunion excursion is situated in the lower part (L28) of the palaeomagnetic sampling trench (**e**).

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.



Extended Data Fig. 10 | See next page for caption.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

Extended Data Fig. 10 | Selected artefacts and fossils from the exploratory trench (subsection KW) in S27 and L28. These are the oldest artefacts discovered at the Shangchen locality during this investigation. **a**, Quartzite core (SC-K4) from the boundary between S27 and L28 at a depth of 60–70 cm below the boundary between cultivated soil and fresh original palaeosol, and 164 cm above the end of the Réunion excursion. See Fig. 4i and Supplementary Tables 6, 7 for details. **b**, **c**, The oldest artefacts SC-K30 and SC-K55 within L28 from the exploratory trench, from 69–114 cm above the end of the Réunion excursion. See Supplementary Table 6 for details. **d**, Quartzite flake tool (SC-K5) from the exploratory trench, from the boundary between S27 and L28 at a depth of about 75 cm below the boundary between cultivated soil and fresh original palaeosol and 149 cm above the end of the Réunion excursion. See Fig. 4h and Supplementary Tables 6, 7 for details. **e**, Flake tool SC-K54 from within the L28 from the exploratory trench. **f**, Mandibular fragment of a small bovid (SC-K60) from the same horizon of S27 as the stone artefacts found during the excavation of S27 and L28. Top image shows the occlusal view of the mandible. **g**, A left mandibular fragment of a large cervid (SC-K10), which was found near and in the same horizon of S27–L28 as the stone artefacts SC-K4 and SC-K5. In the scales, each division represents 1 cm.

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0349-y

Triple oxygen isotope evidence for limited mid-Proterozoic primary productivity

Peter W. Crockford[1,2,3]*, Justin A. Hayles[4,5], Huiming Bao[5,6], Noah J. Planavsky[7], Andrey Bekker[8], Philip W. Fralick[9], Galen P. Halverson[1], Thi Hao Bui[1], Yongbo Peng[5] & Boswell A. Wing[10]

The global biosphere is commonly assumed to have been less productive before the rise of complex eukaryotic ecosystems than it is today[1]. However, direct evidence for this assertion is lacking. Here we present triple oxygen isotope measurements ($\Delta^{17}O$) from sedimentary sulfates from the Sibley basin (Ontario, Canada) dated to about 1.4 billion years ago, which provide evidence for a less productive biosphere in the middle of the Proterozoic eon. We report what are, to our knowledge, the most-negative $\Delta^{17}O$ values (down to −0.88‰) observed in sulfates, except for those from the terminal Cryogenian period[2]. This observation demonstrates that the mid-Proterozoic atmosphere was distinct from what persisted over approximately the past 0.5 billion years, directly reflecting a unique interplay among the atmospheric partial pressures of CO_2 and O_2 and the photosynthetic O_2 flux at this time[3]. Oxygenic gross primary productivity is stoichiometrically related to the photosynthetic O_2 flux to the atmosphere. Under current estimates of mid-Proterozoic atmospheric partial pressure of CO_2 (2–30 times that of pre-anthropogenic levels), our modelling indicates that gross primary productivity was between about 6% and 41% of pre-anthropogenic levels if atmospheric O_2 was between 0.1–1% or 1–10% of pre-anthropogenic levels, respectively. When compared to estimates of Archaean[4–6] and Phanerozoic primary production[7], these model solutions show that an increasingly more productive biosphere accompanied the broad secular pattern of increasing atmospheric O_2 over geologic time[8].

Modern primary producers perform oxygenic photosynthesis, which provides O_2 to the atmosphere and fixes the carbon that fuels heterotrophic consumption in the global biosphere. In the low-O_2 mid-Proterozoic world[8], this oxygenic gross primary productivity (GPP) has been widely assumed to be less than GPP in more recent times[1]. Accordingly, mechanistic explanations for low-O_2 Proterozoic surface environments focus on how to limit the productivity of the Proterozoic biosphere[9–11], despite the lack of direct empirical evidence for lower Proterozoic GPP. Indeed, proposals for sustaining Proterozoic environmental stasis[12–14] would be strengthened if their key component could be validated from the geologic record. Here we examine whether the mid-Proterozoic biosphere was less productive than more recent ones by using a triple oxygen isotope proxy for GPP in sulfate minerals that are dated to about 1.4 billion years ago (Ga).

A record of the productivity of the biosphere is embedded within the mass-independent oxygen isotope fractionation (O-MIF) carried by tropospheric O_2. Stratospheric photochemical reactions preferentially concentrate ^{17}O and ^{18}O in O_3 and leave ^{16}O anomalously enriched in the residual O_2, which therefore has a negative $\Delta^{17}O$ value ($\Delta^{17}O = \delta^{17}O - 0.5305 \times \delta^{18}O$). Isotopic exchange between stratospheric CO_2 and heavy oxygen atoms from O_3 photolysis imparts a positive $\Delta^{17}O$ value to carbon dioxide[15]. Mass balance dictates that O_2 exiting the stratosphere carries a negative $\Delta^{17}O$ value and that the magnitude of this anomaly is proportional to the atmospheric partial pressure of CO_2 (p_{CO_2})[16]. In the troposphere, stratospheric O_2 mixes with the O_2 produced through photosynthesis—which carries ^{16}O, ^{17}O and ^{18}O in isotopically normal proportions—and a near-zero $\Delta^{17}O$ value is set by the isotopic composition of water[17]. The $\Delta^{17}O$ value of tropospheric O_2 reflects the relative contributions of O_2 supplied from the stratosphere versus O_2 derived from photosynthesis[17,18].

This balance is only interpretable if there is enough atmospheric O_2 to sustain a sizable ozone layer: that is, atmospheric partial pressures of O_2 (p_{O_2}) greater than about 0.1% of pre-anthropogenic levels (PAL) (100% PAL = 209,500 p.p.m. O_2)[19]. Although stratospheric oxygen dynamics do not permit the production of O-MIF below this p_{O_2}, such atmospheric anoxia supports the photochemical production of sulfur-isotope mass-independent fractionation (S-MIF). As seen in the Archaean rock record, the presence of S-MIF and absence of O-MIF provides a key metric with which to identify a strictly anoxic atmosphere[20] (Extended Data Table 1). Because photosynthetic carbon fixation is proportional to O_2 production through the photosynthetic quotient, at higher p_{O_2} the $\Delta^{17}O$ anomaly in tropospheric O_2 is a direct measure of GPP[17]. With independent estimates of contemporaneous p_{CO_2} and sufficient p_{O_2}, the $\Delta^{17}O$ value of ancient atmospheric O_2 can constrain ancient GPP[7].

Sedimentary sulfate minerals can preserve a record of ancient atmospheric O_2. Oxidative weathering of sulfide minerals incorporates a sizeable fraction of oxygen from tropospheric O_2 into product sulfate (denoted f_{O_2}). Biologically mediated pyrite oxidation, which dominates abiotic oxidation in natural environments[21], produces a range of f_{O_2} between 8 and 15%[22]. The $\Delta^{17}O$ value of aqueous sulfate is also influenced by the ratio of sulfide to sulfate minerals in the source rocks undergoing oxidative weathering, as well as the intensity and style of microbial sulfur cycling in the aqueous environment[23,24], which erases the $\Delta^{17}O$ signal. For example, modern marine sulfate carries a muted but resolvable negative average $\Delta^{17}O$ value[2] (approximately −0.08‰; Extended Data Table 1) inherited from modern tropospheric O_2 (approximately −0.55‰[25]). Terrestrial sulfate-rich evaporative settings are more conducive than marine settings to the rapid precipitation of gypsum, anhydrite or other sulfate salts[26], and thus both minimize the isotopic consequences of microbial recycling[27] and capture newly formed sulfate with a $\Delta^{17}O$ value that more directly reflects tropospheric O_2.

The Rossport Formation of the Sibley Group in Ontario (Canada) contains lacustrine and sabkha sediments with abundant gypsum layers and nodules. These sulfate-rich sediments are intruded by 1.1-billion-year-old dykes, and have an estimated depositional age of about 1.4 billion years[28] (Supplementary Information). We measured the oxygen ($\delta^{18}O$ and $\Delta^{17}O$) and sulfur ($\delta^{34}S$ and $\Delta^{33}S$) isotope compositions of sulfate that was chemically extracted from 68 Sibley samples. We found $\Delta^{17}O$ values that were more negative than those of modern and Phanerozoic marine sulfates as well as that of modern tropospheric O_2 (Fig. 1), and no evidence of S-MIF (Extended Data Table 1). Our results unambiguously provide a near-direct record of mid-Proterozoic

[1]McGill University, Montreal, Quebec, Canada. [2]Weizmann Institute of Science, Rehovot, Israel. [3]Princeton University, Princeton, NJ, USA. [4]Rice University, Houston, TX, USA. [5]Louisiana State University, Baton Rouge, LA, USA. [6]School of Earth & Space Sciences, Peking University, Beijing, China. [7]Yale University, New Haven, CT, USA. [8]University of California Riverside, Riverside, CA, USA. [9]Lakehead University, Thunder Bay, Ontario, Canada. [10]University of Colorado Boulder, Boulder, CO, USA. *e-mail: peter.crockford@mail.mcgill.ca

© 2018 Springer Nature Limited. All rights reserved.

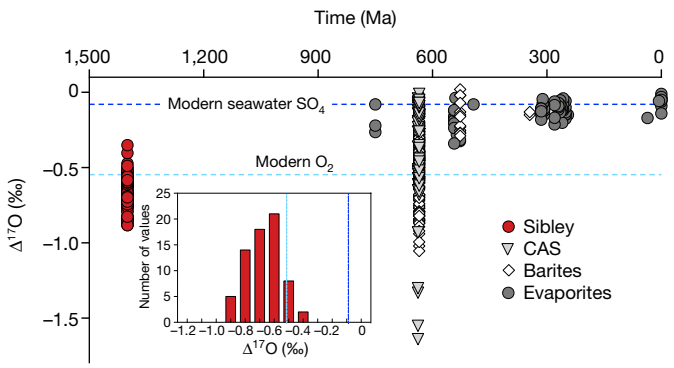


Fig. 1 | Triple oxygen isotope data over the past 1.5 billion years. Compiled $\Delta^{17}O$ data in barites (white diamonds)[2,39–42], carbonate-associated sulfate (CAS) (grey triangles)[29,43] and evaporites (grey or red circles)[2]. Analytical uncertainty on $\Delta^{17}O$ measurements (1σ) is less than 0.05‰. Results from this study ($n = 68$) are compared to average $\Delta^{17}O$ values of modern marine sulfate[2,44] (dark-blue dashed line) and modern tropospheric oxygen[25] (light-blue dashed line). Inset is a histogram of $\Delta^{17}O$ values from the samples analysed here. Ma, million years ago.

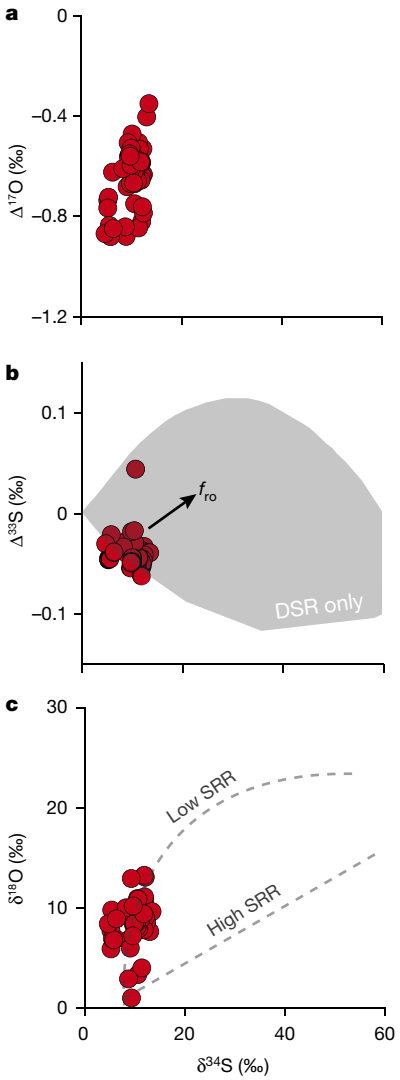


Fig. 2 | Sulfur and oxygen isotope constraints on limited microbial sulfur cycling in the Sibley basin. a, $\delta^{34}S$ and $\Delta^{17}O$ values exhibit a weak positive correlation ($n = 64$; $r^2 = 0.132$ (Pearson's)), which shows that microbial sulfur cycling was not effective in removing anomalous $\Delta^{17}O$ values. **b**, $\delta^{34}S$ and $\Delta^{33}S$ values indicate a sulfur cycle dominated by dissimilatory sulfate reduction, (DSR) (grey field)[30,45] with no evidence for a reoxidative flux into the sulfate pool (f_{ro}, fraction of sulfide that is reoxidized to sulfate). **c**, Steeply aligned $\delta^{34}S$ and $\delta^{18}O$ values indicate relatively slow sulfate reduction rates (SRR)[23]. Total analytical uncertainty on $\delta^{34}S$, $\Delta^{33}S$ and $\delta^{18}O$ measurements is estimated at 0.1, 0.01 and 0.5‰, respectively.

atmospheric O_2 and indicate that the bulk atmosphere has existed in three isotopically distinct states over Earth's history: S-MIF but no O-MIF (Archaean), O-MIF but no S-MIF (Proterozoic), and no substantial O-MIF or S-MIF (Phanerozoic)[2,20] (Extended Data Table 1).

Unlike syn-glacial Cryogenian sulfate (Extended Data Fig. 1), the low $\delta^{34}S$ values, small negative $\Delta^{33}S$ values and high $\delta^{18}O$ values reported here suggest weak microbial sulfur cycling and limited sulfide reoxidation[29,30] in the Sibley basin (Fig. 2). The $\Delta^{17}O$ values of the Sibley Group sulfates, therefore, are primarily modulated by the amount of O_2 incorporated during sulfide oxidation rather than by the weathering and redeposition of pre-existing sulfate minerals or intense reoxidative microbial sulfur cycling (Fig. 2). Considering probable f_{O_2} values for pyrite oxidation (8–15%[22]), the most negative Sibley $\Delta^{17}O$ value (−0.88‰; Extended Data Table 1) indicates that coeval tropospheric O_2 had a $\Delta^{17}O$ value between −11 and −6‰. This range is about 10 to 20 times larger than the $\Delta^{17}O$ of modern tropospheric O_2, which demonstrates that the mid-Proterozoic atmosphere bore little resemblance to the modern atmosphere with respect to the magnitude of its stratospheric and photosynthetic O_2 fluxes as well as the size of the O_2 reservoir in which these fluxes competed.

Although mid-Proterozoic $\Delta^{17}O$ and GPP cannot be calibrated from modern values[2,25], isotope mass balance calculations indicate that the $\Delta^{17}O$ value of tropospheric O_2 reflects three key variables—p_{CO_2}, p_{O_2} and GPP—and that parameters such as troposphere–stratosphere exchange exerted considerably less influence[3]. Under near-modern GPP, as may have characterized the immediate aftermath of the terminal Cryogenian glaciation[31], ultra-high atmospheric p_{CO_2} (more than 350 times PAL; 100% PAL = 280 p.p.m. CO_2) is the only viable way to impart a sizeable negative $\Delta^{17}O$ value to tropospheric oxygen[2,3]. However, in conditions of lower p_{CO_2} similarly negative $\Delta^{17}O$ values can be generated if GPP is greatly diminished, which effectively lengthens the residence time of photosynthetic O_2 in the atmosphere[3]. We used this isotope mass-balance approach[3] together with independent estimates for mid-Proterozoic levels of p_{CO_2} and p_{O_2} to show that the $\Delta^{17}O$ values of the Sibley sulfate provide empirical evidence that limited mid-Proterozoic primary productivity was a biogeochemical reality.

Estimates of mid-Proterozoic p_{CO_2} from geochemical proxies and climate models suggest p_{CO_2} between 2× and 30× PAL at 1.4 Ga (Extended Data Fig. 2). Mid-Proterozoic O_2 estimates span a wider relative range but fall between 0.1 and 10% of PAL (Extended Data Fig. 2). We take an agnostic perspective of the techniques for estimating p_{O_2} and p_{CO_2} and use uniform frequency distributions within these ranges as the basis for a Monte Carlo resampling of mid-Proterozoic O_2 and CO_2 levels. Along with a resampling of other control parameters in the isotope mass-balance model (Extended Data Fig. 3), this enabled us to turn our measured $\Delta^{17}O$ distribution (Fig. 1) into bootstrap estimates for mid-Proterozoic GPP (Methods). The Monte Carlo estimates confirm that it is very likely that the mid-Proterozoic biosphere was less productive than the present biosphere, with >93% of the GPP estimates being less than pre-anthropogenic levels (Fig. 3).

There is a small tail of relative GPP values above 100% PAL that require sulfate $\Delta^{17}O$ approaching zero, a f_{O_2} near the maximum probable value and—importantly—p_{O_2} close to 10% PAL (Extended Data Fig. 4). Although permitted by our analysis, these extreme circumstances were unlikely to have been persistent on the mid-Proterozoic Earth, if they were present at all. Mid-Proterozoic GPP estimates are always strongly influenced by p_{O_2} (Extended Data Fig. 4b), which suggests that the bimodal structure of the GPP probability density (Fig. 3) is a consequence of the wide range of p_{O_2} values that we used in the Monte Carlo calculations. Under more restricted ranges of p_{O_2}, the

© 2018 Springer Nature Limited. All rights reserved.

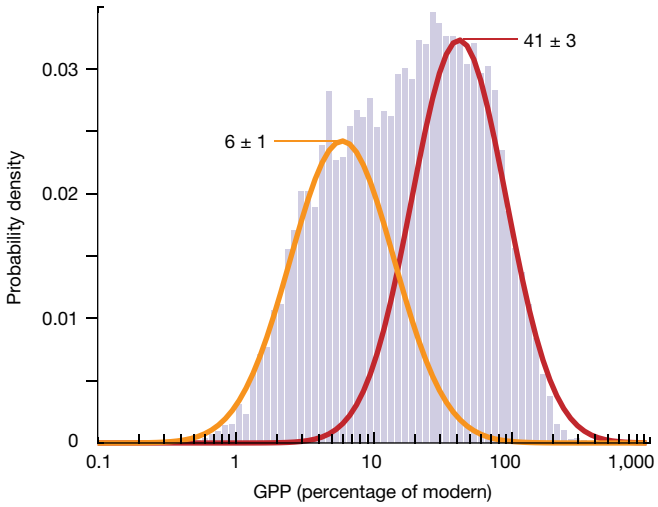


Fig. 3 | Empirical probability densities of mid-Proterozoic GPP. Monte Carlo solutions are based on the measured distribution of $\Delta^{17}O$ values of the approximately 1.4-billion-year-old Sibley sulfates, and take into account uncertainties in the control parameters of the isotope mass-balance model. The grey histogram represents the full suite of estimates and is well-approximated as the sum of a pair of Gaussian probability density functions (root mean squared error = 0.001; Methods). The red line represents the probability density function for GPP solutions when p_{O_2} levels are 1–10% PAL, which is characterized by a mean GPP value of 41 ± 3% PAL (uncertainties are 95% confidence intervals on the mean). The orange line represents the probability density function for GPP solutions when p_{O_2} levels are 0.1–1% PAL, which is characterized by a mean GPP value of 6 ± 1% PAL.

bimodality disappears and the estimated probability densities for GPP take on a normal form (Extended Data Fig. 5, Methods). To explore this behaviour, we fit the full suite of GPP results to a pair of Gaussian probability density functions, one with a mean GPP of 6 ± 1% PAL that is representative of low O_2 conditions (0.1–1% PAL) and another with a mean GPP of 41 ± 3% PAL (quoted uncertainties are 95% confidence intervals on the mean; Fig. 3) that is representative of high O_2 conditions (1–10% PAL). The lower GPP state is consistent with lower mean p_{O_2} (about 0.3% PAL), and the higher GPP state corresponds to higher mean p_{O_2} (about 2% PAL; Extended Data Fig. 5). The trajectory of these estimates reflects the positive covariation between GPP and p_{O_2} that scales across a range of GPP and p_{O_2} values (Extended Data Fig. 4b).

Approximately 700–800 million years after the deposition of the Sibley sulfates, there was a major ecological shift in the marine realm that saw a cyanobacterial phytoplankton population replaced by one dominated by algae[32]. This ecological transition has previously been linked to a rise in nutrient levels[12], consistent with eukaryotic predominance in more nutrient-rich regions of the modern oceans[33]. In light of the timing of this transition and the marked similarity of our higher GPP estimate to algal-dominated primary production in today's ocean (about 45% of the primary production of the modern biosphere[34]), we suggest that our lower GPP estimate (and, as a result, a lower p_{O_2}) is more likely to reflect the mid-Proterozoic world. Enhanced regenerated production in a bacteria-dominated, oligotrophic marine ecosystem[35,36] could have supported higher overall GPP in nutrient-limited Proterozoic oceans[1]. However, there is evidence that intense heterotrophic recycling of carbon appeared only in planktonic ecosystems in the latest Proterozoic[32,37,38], providing further support for our lower GPP–p_{O_2} estimate. Although reduced Proterozoic GPP is typically linked to the need to throttle O_2 generation on geologic timescales via diminished export production and organic carbon burial[9,11,12], the GPP–p_{O_2} covariation described here highlights the fact that diminished O_2 production may have also directly led to reduced O_2 levels on the Proterozoic Earth. Along with Archaean[4–6] and Phanerozoic[7] GPP estimates that bookend our inferred Proterozoic GPP (about 6% PAL), this relationship suggests that primary production has progressively increased in concert with the broad two-step history of O_2 in Earth's atmosphere[8].

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0349-y.

Received: 20 March 2017; Accepted: 1 June 2018;
Published online 18 July 2018.

1. Anbar, A. D. & Knoll, A. H. Proterozoic ocean chemistry and evolution: a bioinorganic bridge? *Science* **297**, 1137–1142 (2002).
2. Bao, H., Lyons, J. R. & Zhou, C. Triple oxygen isotope evidence for elevated CO_2 levels after a Neoproterozoic glaciation. *Nature* **453**, 504–506 (2008).
3. Cao, X. & Bao, H. Dynamic model constraints on oxygen-17 depletion in atmospheric O_2 after a snowball Earth. *Proc. Natl Acad. Sci. USA* **110**, 14546–14550 (2013).
4. Kharecha, P., Kasting, J. & Siefert, J. A coupled atmosphere–ecosystem model of the early Archean Earth. *Geobiology* **3**, 53–76 (2005).
5. Canfield, D. E., Rosing, M. T. & Bjerrum, C. Early anaerobic metabolisms. *Phil. Trans. R. Soc. Lond. B* **361**, 1819–1836 (2006).
6. Ward, L. M., Kirschvink, J. L. & Fischer, W. W. Timescales of oxygenation following the evolution of oxygenic photosynthesis. *Orig. Life Evol. Biosph.* **46**, 51–65 (2016).
7. Wing, B. A. A cold, hard look at ancient oxygen. *Proc. Natl Acad. Sci. USA* **110**, 14514–14515 (2013).
8. Lyons, T. W., Reinhard, C. T. & Planavsky, N. J. The rise of oxygen in Earth's early ocean and atmosphere. *Nature* **506**, 307–315 (2014).
9. Laakso, T. A. & Schrag, D. P. Regulation of atmospheric oxygen during the Proterozoic. *Earth Planet. Sci. Lett.* **388**, 81–91 (2014).
10. Sánchez-Baracaldo, P., Ridgwell, A. & Raven, J. A. A Neoproterozoic transition in the marine nitrogen cycle. *Curr. Biol.* **24**, 652–657 (2014).
11. Derry, L. A. Causes and consequences of mid-Proterozoic anoxia. *Geophys. Res. Lett.* **42**, 8538–8546 (2015).
12. Reinhard, C. T. et al. Evolution of the global phosphorus cycle. *Nature* **541**, 386–389 (2017).
13. Koehler, M. C. et al. Spatial and temporal trends in Precambrian nitrogen cycling: a Mesoproterozoic offshore nitrate minimum. *Geochim. Cosmochim. Acta* **198**, 315–337 (2017).
14. Buick, R., Des Marais, D. J. & Knoll, A. H. Stable isotopic compositions of carbonates from the Mesoproterozoic Bangemall Group, northwestern Australia. *Chem. Geol.* **123**, 153–171 (1995).
15. Wen, J. & Thiemens, M. H. Multi-isotope study of the O (^1D)+ CO_2 exchange and stratospheric consequences. *J. Geophys. Res.* **98**, 12801–12808 (1993).
16. Yung, Y. L., DeMore, W. B. & Pinto, J. P. Isotopic exchange between carbon dioxide and ozone via O(^{1}D) in the stratosphere. *Geophys. Res. Lett.* **18**, 13–16 (1991).
17. Luz, B., Barkan, E., Bender, M. L., Thiemens, M. H. & Boering, K. A. Triple-isotope composition of atmospheric oxygen as a tracer of biosphere productivity. *Nature* **400**, 547–550 (1999).
18. Bender, M., Sowers, T. & Labeyrie, L. The Dole effect and its variations during the last 130,000 years as measured in the Vostok ice core. *Glob. Biogeochem. Cycles* **8**, 363–376 (1994).
19. Segura, A. et al. Ozone concentrations and ultraviolet fluxes on Earth-like planets around other stars. *Astrobiology* **3**, 689–708 (2003).
20. Bao, H., Rumble, D. III & Lowe, D. R. The five stable isotope compositions of Fig Tree barites: implications on sulfur cycle in ca. 3.2 Ga oceans. *Geochim. Cosmochim. Acta* **71**, 4868–4879 (2007).
21. Percak-Dennett, E. et al. Microbial acceleration of aerobic pyrite oxidation at circumneutral pH. *Geobiology* **15**, 690–703 (2017).
22. Balci, N., Shanks, W. C. III, Mayer, B. & Mandernack, K. W. Oxygen and sulfur isotope systematics of sulfate produced by bacterial and abiotic oxidation of pyrite. *Geochim. Cosmochim. Acta* **71**, 3796–3811 (2007).
23. Antler, G., Turchyn, A. V., Rennie, V., Herut, B. & Sivan, O. Coupled sulfur and oxygen isotope insight into bacterial sulfate reduction in the natural environment. *Geochim. Cosmochim. Acta* **118**, 98–117 (2013).
24. Pellerin, A. et al. Mass-dependent sulfur isotope fractionation during reoxidative sulfur cycling: a case study from Mangrove Lake, Bermuda. *Geochim. Cosmochim. Acta* **149**, 152–164 (2015).
25. Pack, A. et al. Tracing the oxygen isotope composition of the upper Earth's atmosphere using cosmic spherules. *Nat. Comm.* **8**, 15702 (2017).
26. Hardie, L. A. The origin of the recent non-marine evaporite deposit of Saline Valley, Inyo County, California. *Geochim. Cosmochim. Acta* **32**, 1279–1301 (1968).
27. Ryu, J.-H., Zierenberg, R. A., Dahlgren, R. A. & Gao, S. Sulfur biogeochemistry and isotopic fractionation in shallow groundwater and sediments of Owens Dry Lake, California. *Chem. Geol.* **229**, 257–272 (2006).

© 2018 Springer Nature Limited. All rights reserved.

28. Rogala, B., Fralick, P. W., Heaman, L. M. & Metsaranta, R. Lithostratigraphy and chemostratigraphy of the Mesoproterozoic Sibley Group, northwestern Ontario, Canada. *Can. J. Earth Sci.* **44**, 1131–1149 (2007).
29. Bao, H., Fairchild, I. J., Wynn, P. M. & Spötl, C. Stretching the envelope of past surface environments: Neoproterozoic glacial lakes from Svalbard. *Science* **323**, 119–122 (2009).
30. Johnston, D. T. et al. Active microbial sulfur disproportionation in the Mesoproterozoic. *Science* **310**, 1477–1479 (2005).
31. Kunzmann, M. et al. Zn isotope evidence for immediate resumption of primary productivity after snowball Earth. *Geology* **41**, 27–30 (2013).
32. Brocks, J. J. et al. The rise of algae in Cryogenian oceans and the emergence of animals. *Nature* **548**, 578–581 (2017).
33. Ward, B. A., Dutkiewicz, S. & Follows, M. J. Modelling spatial and temporal patterns in size-structured marine plankton communities: top–down and bottom–up controls. *J. Plankton Res.* **36**, 31–47 (2013).
34. Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. Primary production of the biosphere: integrating terrestrial and oceanic components. *Science* **281**, 237–240 (1998).
35. Eppley, R. W. & Peterson, B. J. Particulate organic matter flux and planktonic new production in the deep ocean. *Nature* **282**, 677–680 (1979).
36. Close, H. G. et al. Export of submicron particulate organic matter to mesopelagic depth in an oligotrophic gyre. *Proc. Natl Acad. Sci. USA* **110**, 12565–12570 (2013).
37. Logan, G. A., Hayes, J. M., Hieshima, G. B. & Summons, R. E. Terminal Proterozoic reorganization of biogeochemical cycles. *Nature* **376**, 53–56 (1995).
38. Close, H. G., Bovee, R. & Pearson, A. Inverse carbon isotope patterns of lipids and kerogen record heterogeneous primary biomass. *Geobiology* **9**, 250–265 (2011).
39. Peng, Y., Bao, H., Zhou, C. & Yuan, X. ^{17}O-depleted barite from two Marinoan cap dolostone sections, South China. *Earth Planet. Sci. Lett.* **305**, 21–31 (2011).
40. Killingsworth, B. A., Hayles, J. A., Zhou, C. & Bao, H. Sedimentary constraints on the duration of the Marinoan Oxygen-17 Depletion (MOSD) event. *Proc. Natl Acad. Sci. USA* **110**, 17686–17690 (2013).
41. Crockford, P. W. et al. Linking paleocontinents through triple oxygen isotope anomalies. *Geology* **46**, 179–182 (2017).
42. Crockford, P. W. et al. Triple oxygen and multiple sulfur isotope constraints on the evolution of the post-Marinoan sulfur cycle. *Earth Planet. Sci. Lett.* **435**, 74–83 (2016).
43. Bao, H., Chen, Z.-Q. & Zhou, C. An ^{17}O record of late Neoproterozoic glaciation in the Kimberley region, Western Australia. *Precambr. Res.* **216–219**, 152–161 (2012).
44. Cowie, B. R. & Johnston, D. T. High-precision measurement and standard calibration of triple oxygen isotopic compositions (δ^{18}O, Δ'^{17}O) of sulfate by F_2 laser fluorination. *Chem. Geol.* **440**, 50–59 (2016).
45. Sim, M. S., Bosak, T. & Ono, S. Large sulfur isotope fractionation does not require disproportionation. *Science* **333**, 74–77 (2011).

Acknowledgements P.W.C. acknowledges funding from a Natural Sciences and Engineering Research Council of Canada (NSERC) PGS-D grant. This research was supported by funding from NSERC, the Fonds de Recherche du Québec–Nature et Technologies, and the University of Colorado Boulder (B.A.W.). H.B. acknowledges funding from the strategic priority research program (B) of CAS (XDB18010104). We thank E. Wolf, N. Cowan, S. Becker, Y. Slichter and L. Derry for discussions on the assumptions, methods and implications of our study.

Reviewer information *Nature* thanks B. Passey and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions P.W.C. and B.A.W. designed research. N.J.P., A.B. and P.W.F. provided samples. P.W.C., J.A.H., Y.P. and T.H.B. performed isotopic analyses. P.W.C. and B.A.W. conducted modelling. P.W.C. wrote the manuscript with contributions from B.A.W., N.J.P. and A.B., and input from all co-authors.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0349-y.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0349-y.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to P.W.C.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Sample information. *Sample locations.* Sediments of the Sibley Group were deposited over an approximately 70,000 km^2 area that extends in to northern Lake Superior, and to the margins of Lake Nipigon in Ontario, Canada (Extended Data Fig. 6). While the surface expression of the Sibley Group is typically poor, extensive mineral exploration in the area has provided kilometres of drill core. For this study, 68 samples were collected from four different drill cores: NI-92-7 (UTM east 353850, north 5443000), NB-97-2 (UTM east 426990 north 5416241), NB-97-4 (UTM east 425430 north 5410540) and WP-07-03 (UTM east 379602 north 5422047).

Age constraints. Maximum age constraints for the entire Sibley Group are provided by rhyolites that unconformably underlie the Sibley Group with a U–Pb date of 1,530.5 ± 6.2 million years ago (Ma)[46]. A minimum age is constrained by cross-cutting diabase sills dated at 1,109.7 ± 2 Ma[46]. The only age measured directly on Sibley Group material is provided by a whole-rock Rb–Sr isochron that gave a calculated minimum date of 1,339 ± 33 Ma for the Kama Hill Formation shales[47]. Palaeomagnetic pole positions place the Sibley Group[48] on the apparent polar wander path of the Belt Supergroup at about 1.4 Ga[49]. In light of these constraints and the stratigraphic location of the Rossport Formation below the Kama Hill Formation (see below), we consider the approximate age of the material analysed for this study to be 1.4 Ga.

Depositional setting. Gypsum nodules used in this study were taken from the Firehill Member of the Rossport Formation, which was deposited after the fluvial to near-shore lacustrine sandstones and conglomerates of the Pass Lake Formation and before the saline mud-flats of the Kama Hill Formation[28]. The Rossport Formation is subdivided into three members beginning with the Channel Island Member at the base[50]. The Channel Island Member is characterized by cyclic siltstone–dolostone couplets with evaporite minerals preferentially being deposited in dolostone layers as nodules and bladed crystals[51]. The Channel Island Member contains massive beds of sandstone near its top and is overlain by the Middlebrun Bay Member[50]. This 0.5–3-m thick, stromatolitic dolostone-chert represents the shoreline of the saline lake[28]. The Middlebrun Bay Member is overlain by the Firehill Member, which consists of massive to finely laminated red siltstones with sporadic intraformational mass-flow conglomerate beds. Deposition occurred on mud- and sand-flats with a near-surface, saline water table[28,51]. Sulfate minerals throughout the Firehill Member have been observed to occur as nodules, bladed crystals, veins, cements and detrital grains; however, the sample set in this study consisted of only nodules. The Firehill Member is terminated by the appearance of ripple marks and hummocky cross-bedded sandstones and mudstones of the Kama Hill Formation, marking the transgression of a large water mass. Together these observations have shaped the view that the Pass Lake and Rossport Formations of the Sibley Group were deposited in predominantly fluvio-lacustrine and sabkha settings, although additional work is needed to constrain the influence of marine waters on the system.

Isotopic measurements. *$\Delta^{17}O$ measurements.* We followed a previously detailed analytical procedure[2,42]. Drill core samples with abundant gypsum nodules were selected for oxygen isotope analyses (Extended Data Table 1). After a thin layer was milled from the outer surface to avoid potential modern contamination, samples were drilled to collect about 30 mg of material for pre-treatment. Drilled powders were first dissolved into a 0.1 M sodium hydroxide–0.05 M diethylenetriaminepentaacetic acid (DTPA) solution to extract sulfate into solution and remove any non-sulfate oxygen-bearing species[52]. The extracted sulfate samples were then reprecipitated at 80 °C by acidifying with double-distilled 6 M hydrochloric acid followed by the addition of drops of concentrated barium chloride solution. This dissolution and reprecipitation was repeated to further eliminate possible contaminations. Approximately 10 mg of purified $BaSO_4$ was then loaded onto a 316L stainless steel plate and placed under a bromine pentafluoride (BrF_5) atmosphere for 12 h to eliminate any water absorbed by the samples. Molecular oxygen was generated from the samples using a CO_2-laser fluorination system. Although approximately 25 μmol of O_2 gas was generated (25–35% yield) for each sample, this process is not observed to induce any significant isotopic fractionation for $\Delta^{17}O$ values[53]. Molecular O_2 produced during the fluorination process was then taken through a number of cryo-focusing steps to remove condensable gases, followed by collection onto mol sieve 5A at −196 °C before expansion into the sample bellows of a Thermo MAT 253 mass spectrometer for dual-inlet analysis.

$\Delta^{17}O$ values are expressed arithmetically as $\Delta^{17}O = \delta^{17}O - 0.5305 \times \delta^{18}O$ (0.5305 represents a high temperature limit for ^{16}O–^{17}O–^{18}O fractionation[54–56]) and $\delta^iO = ([^iR_{sample}/^iR_{V\text{-}SMOW}] - 1) \times 1{,}000$, in which ^iR represents the ratio of ^iO (i = 17 or 18) to ^{16}O, and calculated values are reported as parts per thousand (‰) on the Vienna Standard Mean Ocean Water (V-SMOW) scale. We used an arithmetic definition for $\Delta^{17}O$ instead of a logarithmic definition—for example, $\Delta^{17}O = \ln(\delta^{17}O + 1) - 0.5305 \times \ln(\delta^{18}O + 1)$—to enable direct comparison between measured $\Delta^{17}O$ values and $\Delta^{17}O$ values calculated from the isotope mass-balance model. Although a logarithmic definition for $\Delta^{17}O$ is commonly used owing to a weaker dependence on the isotopic composition of reference material[57,58], differences between arithmetic and logarithmic $\Delta^{17}O$ values calculated in this study were typically in the order of 0.001‰. This deviation is much less than the uncertainty on $\Delta^{17}O$ analyses (1σ = 0.05‰), estimated based on multiple measurements of the same $BaSO_4$ calibrated against UWG-2 (taken to have $\delta^{18}O$ = 5.80‰ and $\delta^{17}O$ = 3.016‰ on the V-SMOW scale; Extended Data Table 1, Extended Data Fig. 1).

$\delta^{18}O$ measurements. Because the low yield of laser-fluorination techniques leads to a significant fractionation in the measured $\delta^{18}O$ value of sulfates, we combusted sulfate samples and measured the major oxygen isotope composition on the resulting carbon monoxide (CO)[44]. Measurements for $\delta^{18}O$ values were made on the same aliquots of sample used for $\Delta^{17}O$ analysis that underwent the DTPA reprecipitation treatment. Analyses were performed using a Temperature Conversion Elemental Analyzer connected to a Conflo-III and a Thermo MAT 253 mass spectrometer in continuous-flow mode. The estimated total analytical error for $\delta^{18}O$ analyses is estimated to be less than 0.5‰ from repeated measurements of laboratory standards (USGS34 $\delta^{18}O$ = −27.80‰; IAEA-NO3 $\delta^{18}O$ = 25.6‰; USGS35 $\delta^{18}O$ = 56.80‰)[42].

$\delta^{34}S$ and $\Delta^{33}S$ measurements. Sulfur isotope measurements were made using purified $BaSO_4$ produced from the DTPA re-precipitation treatment for oxygen isotope analysis. Approximately 10 mg of barite powder was reacted with 15 ml of Thode reduction solution at 100 °C for at least 2 h[59]. Powders reacted completely to produce H_2S that was carried through a condenser in a N_2 gas stream and bubbled into a 0.4 M zinc acetate solution in which the H_2S was quantitatively captured as ZnS. Samples were then reacted with ~5–10 drops of 0.2 M $AgNO_3$ solution to convert ZnS to Ag_2S. The Ag_2S was then filtered, collected and dried for 12 h. Approximately 3 mg aliquots of dried Ag_2S were then loaded into nickel bombs and heated to 250 °C for 12 h under a fluorine gas atmosphere to generate SF_6 gas for analysis. The resulting SF_6 was purified cryogenically and chromatographically before isotopic analysis on a Thermo MAT 253 mass spectrometer in dual-inlet mode.

Sulfur isotope compositions are expressed as $\delta^iS = ([^iR_{sample}/^iR_{V\text{-}CDT}] - 1) \times 1{,}000$, in which ^iR represents the ratio of ^iS (i = 33 or 34) to ^{32}S, and V-CDT indicates the Vienna Canon Diablo Troilite scale. $\Delta^{33}S$ values are expressed as $\Delta^{33}S = \delta^{33}S - 1000 \times ([1 + (\delta^{34}S/1{,}000)]^{0.515} - 1)$.

Results were normalized against repeat analyses of international reference material IAEA-S1, which we take to have $\delta^{34}S$ = −0.3‰ and $\Delta^{33}S$ = −0.061‰ on the V-CDT scale. The estimated (1σ) total analytical uncertainty on the entire procedure is estimated to be better than 0.1‰ for $\delta^{34}S$ and 0.01‰ for $\Delta^{33}S$ (Extended Data Table 1).

Model description. We adapted an analytical solution to a steady-state model[3] for the $\Delta^{17}O$ value of tropospheric O_2 that allows for variation in p_{O_2}, p_{CO_2}, stratosphere–troposphere transport and stratospheric O_2–O_3–CO_2 photochemistry[55,60] (Extended Data Table 2). As described below, we made three primary changes to this model. First, we cast the model solution into a form that incorporated p_{O_2}, p_{CO_2} and GPP estimates relative to a set of reference values for PAL (Extended Data Table 2). Second, we rearranged the model solution to solve GPP as a function of measured $\Delta^{17}O$ values and model parameters. Third, we inverted the model solution to solve p_{O_2} as a function of measured $\Delta^{17}O$ values, other model parameters and inferred GPP. We present the equations used to enact these changes in the Supplementary Information and describe the constraints that we applied to the various model parameters below. Finally, we outline our Monte Carlo approach for estimating probability densities for GPP and p_{O_2}.

Constraints on model parameters. p_{CO_2}. One-dimensional radiative convective models have been used to estimate p_{CO_2} levels required to maintain Earth surface temperatures of 273 K and 288 K, placing broad upper and lower limits on p_{CO_2} at 1.4 Ga of 1 and 100 PAL, respectively[61] (Extended Data Fig. 2). A further accounting of transport process through extrapolation of results from the CAM3 general circulation model refines these values considerably[62]. For example, to maintain an average surface temperature of 288 K under a pure CO_2 atmosphere an upper estimate of p_{CO_2} at 1.4 Ga is calculated to be 30 PAL, and an atmosphere including 1 PAL N_2 and 10^{-4} bar of methane provides a lower bound at 5 PAL (Extended Data Fig. 2). These ranges are consistent with results from the COPSE Earth System model that put forward a range of 8–20 PAL[63] for p_{CO_2} at 1.4 Ga (Extended Data Fig. 2).

Geochemical approaches have also been used to constrain Proterozoic CO_2 levels. A previous study attempted to constrain p_{CO_2} at 1.4. Ga by relating carbon isotope fractionations between organic matter in microfossils and surrounding carbonate to extracellular CO_2 levels[64]. Such estimates remain poorly calibrated in the laboratory, especially at high CO_2 levels, so we do not include the suggested 10–200 PAL range put forward through this approach. Quantification of silicate weathering in response to different levels of p_{CO_2} in profiles of 1.8-billion-year-old (p_{CO_2} = 45 PAL) and 1.1-billion-year-old (p_{CO_2} = 1 PAL) palaeosols suggests p_{CO_2} levels were about 2–20 PAL when extrapolated to 1.4 Ga[65]. A lower bound

© 2018 Springer Nature Limited. All rights reserved.

of 2 PAL approximates the CO_2 threshold (350–550 p.p.m.) thought to maintain an ice-sheet-free Palaeogene Earth[66] (Extended Data Fig. 2). Distilling results from these different approaches and considering the robustness of the different results, we set a uniform uncertainty envelope around CO_2 levels at 1.4 Ga of 2–30 PAL[62,65] (Extended Data Fig. 2).

p_{O_2}. The presence of $\Delta^{17}O$ anomalies requires a threshold amount of atmospheric oxygen to allow for the inception of photochemical reactions involving ozone. Calculations suggest this threshold exists at a p_{O_2} value of 0.001 PAL[19,67] (Extended Data Fig. 2), well above the levels calculated for the removal of mass-independent sulfur isotope anomalies[68]. It is for this reason that Archaean samples[69] (Extended Data Table 1) do not bear anomalous $\Delta^{17}O$ signatures[20] even though Archaean p_{CO_2} concentrations were probably at much higher levels than the modern environment.

Studies based on mineral stability observed through iron loss within palaeosols estimated that mid-Proterozoic O_2 levels were greater than 0.01 PAL[70]. Recent work has examined the kinetics involved in oxidizing terrestrial Mn or Fe and tracked this through the isotopic composition of chromium (Cr) through the mid-Proterozoic sedimentary record. These Cr-isotope-based estimates provide a threshold estimate of mid-Proterozoic p_{O_2} at <0.01–0.001 PAL, with evidence for p_{O_2} levels above this only appearing after 1.2 – 0.8 Ga[71–73] (Extended Data Fig. 2). Palaeosol and Cr-isotope-based estimates contrast with those suggested through trace metal enrichments and biomarkers in 1.4-billion-year-old shales that have been used to argue for p_{O_2} levels >0.04 PAL[74,75]. These estimates are consistent with a broad range of values implied by tracking Zn/Fe ratios of carbonates over the past 3.5 billion years of Earth history that place p_{O_2} at less than 0.06 PAL over the mid-Proterozoic[76].

Modelling studies have added to this debate by suggesting that mid-Proterozoic p_{O_2} must be much less than 0.4 PAL[77] to be consistent with no evidence for persistent fully oxygenated oceans in preserved marine sediments[78,79] or 0.1 PAL to be consistent with the stability of the stable carbon isotope record[80]. Biological modelling constraints have been provided by calculating O_2 requirements for *Grypiania spiralis*, and were suggested to be between 0.01–0.1 PAL, placing a lower limit on p_{O_2} upon the emergence of these fossils in the geologic record by 1.4 Ga[81]. However, estimates using this logic have been brought down considerably in recent years to about 0.0015 PAL from calculations based on minimum requirements of simple bilaterians and about 0.004 PAL from experimental results on sponges grown in the laboratory[82,83] (Extended Data Fig. 2).

All of these studies speak to reduced p_{O_2} in the mid-Proterozoic compared to more recent Earth history (Extended Data Fig. 2), but to date there is no clear consensus. With this in mind, we remain largely agnostic to the techniques used to derive previous estimates and apply a uniform distribution between a lower limit of 0.001 PAL (0.1% PAL) and an upper limit of 0.1 PAL (10% PAL) as our initial estimate for mid-Proterozoic p_{O_2} (Extended Data Table 2, Extended Data Fig. 3).

f_{O2}. Two laboratory studies[22,84] have attempted to quantify different pathways of pyrite oxidation and the proportion of atmospheric O_2 that is incorporated into product sulfate (f_{O_2}). In experiments with *Acidithiobacillus ferrooxidans* it was determined that between 8 and 15% of oxygen in product sulfate was from atmospheric oxygen[22]. In abiotic experiments it was determined that 13% of oxygen in product sulfate was from atmospheric oxygen[22]. Both experiments were conducted at low pH values of 2.2–3[22]. In a second study, abiotic experiments were conducted over a much broader pH range (2–11) and O_2 incorporation into sulfate during pyrite oxidation was determined using both major ($\delta^{18}O$) and minor ($\Delta^{17}O$) oxygen isotopes[84]. In these experiments it was determined that 21–34% of oxygen in sulfate was sourced from atmospheric oxygen[84]. Although it remains difficult to determine the exact proportion of oxygen within sulfate that is sourced from H_2O and O_2, these previous studies provide a conservative range of f_{O_2} of 8–34%. When considering the differences in kinetics between abiotic and biotic pyrite oxidation in modern settings, however, this range can be refined. Experiments and natural observations have shown that biological mediation can markedly increase pyrite oxidation kinetics and therefore is more likely to dominate natural surface environments both at present and probably in the past[21,84]. Therefore, in this study we assume a uniform distribution of possible f_{O_2} values between 8 and 15% (Extended Data Table 2, Extended Data Fig. 3).

Oxygen isotope fractionation in O_2–O_3–CO_2 photochemical reaction networks. We follow a previously published approach[3] and use previously published results of photochemical experiments[60] to constrain the parameters: ρ_0, X_l, X_h, ${}^{17}m^{strat}_{CO_2/O_2}$ and ${}^{17}b^{strat}_{CO_2/O_2}$. Published fits to the experimental results for $\delta^{18}O^{strat}_{CO_2/O_2}$ as a function of p_{O_2}/p_{CO_2} ratios constrain ρ_0 to 1.230 ± 0.325, X_l to 64.0 ± 5.5‰ and X_h to 146 ± 2‰[60] (all uncertainties 1σ; Extended Data Table 2). We used these values and their uncertainties to define Gaussian probability density functions for ρ_0, X_l and X_h (Extended Data Fig. 3).

We made a linear least-squares fit to the $\delta^{18}O^{strat}_{CO_2/O_2}$ and $\delta^{17}O^{strat}_{CO_2/O_2}$ measurements from a previously published study[60] (Extended Data Table 1) and used this fit to derive a linear relationship between $\delta^{17}O^{strat}_{CO_2/O_2}$ and $\delta^{18}O^{strat}_{CO_2/O_2}$. The coefficients of the resulting relationship are: ${}^{17}m^{strat}_{CO_2/O_2} = 0.5167 \pm 0.0336$ and ${}^{17}b^{strat}_{CO_2/O_2} = -8.052 \pm 3.673$ (all uncertainties 1σ; Extended Data Table 2). We used these values and their uncertainties to define Gaussian probability density functions for ${}^{17}m^{strat}_{CO_2/O_2}$ and ${}^{17}b^{strat}_{CO_2/O_2}$ (Extended Data Fig. 3). The coefficients derived here differ slightly from the previously published approach[3] owing to the fact that we define $\Delta^{17}O$ values with a slope of 0.5305[56,85] rather than 0.52.

Gamma (γ). We use an estimate of the mass of the modern atmosphere (5.148×10^{18} kg)[86] and a recent determination of the stratosphere–troposphere mass flux (2.194×10^{17} kg yr^{-1})[87] to derive an updated estimate for the stratosphere–troposphere exchange rate in the modern atmosphere (0.0426 yr^{-1}; Extended Data Table 2). It is not well-known how stratosphere–troposphere mass transfer will change under atmospheric states different from the modern state. Previous work has explored how the Brewer–Dobson circulation may change with elevated CO_2 levels, and suggests that increasing p_{CO_2} by a factor of two will increase mass flux across the tropopause by about 20% with further CO_2 increases only slightly affecting stratosphere–troposphere exchange[88]. As we lack a full quantitative exploration of the effect of mid-Proterozoic atmospheric state on γ, we take a statistical approach here and allow γ to vary by about 50% according to a Gaussian probability function defined by a 1σ value of 0.01 (Extended Data Table 2, Extended Data Fig. 3).

Theta (θ). We use the calculation procedure described above to define a value for the fraction of stratospheric O_2 taking part in the stratospheric O_2–O_3–CO_2 reaction network under pre-anthropogenic levels ($p_{O_2} = 1$ PAL; $p_{CO_2} = 1$ PAL; GPP = 1 PAL). Within this calculation framework, a value of θ of 0.1156 reproduces the $\Delta^{17}O$ of modern tropospheric O_2 (−0.546‰ under the definition of $\Delta^{17}O$ used here)[25]. Although it is unknown how this value will change under different atmospheric states, we assume that θ is fixed by the development of an ozone layer and hold it constant throughout the calculations presented here.

Model calculations and sensitivity. We estimated probability densities for mid-Proterozoic GPP by repeating the calculation procedure described above 10,000 times, each time drawing a single estimate for each model parameter according to the distributions shown in Extended Data Fig. 3 and a single estimate for the $\Delta^{17}O$ of mid-Proterozoic sulfate randomly from the dataset measured here (Fig. 1, Extended Data Table 1, Extended Data Fig. 7). A realization of the GPP probability density from this Monte Carlo resampling and calculation procedure is shown in Fig. 3 and archived in Supplementary Table 1. We also plot individual GPP results against various control parameters of the model in Extended Data Fig. 4. Although we permitted potential variability in $\Delta^{17}O$ due to local processes (for example, changes in sulfide oxidation through variable f_{O_2}) that may have varied stratigraphically (Extended Data Fig. 7), as well as global processes (for example, changes to stratospheric oxygen dynamics through a variable stratosphere–troposphere exchange rate), these plots show that GPP is most sensitive to variations in p_{O_2}, with a near log-linear relationship between assumed p_{O_2} and calculated GPP (Extended Data Fig. 4b). Much weaker dependencies on f_{O_2} and $\Delta^{17}O_{sulfate}$ are apparent. Although a small percentage of the calculated GPP values are greater than 1 PAL, the sensitivity exercise illustrates that these values require a combination of high p_{O_2}, small $\Delta^{17}O_{sulfate}$ and large f_{O_2} (Extended Data Fig. 4). This combination of extreme conditions is unlikely to have characterized the mid-Proterozoic Earth, given the probability distributions derived here (Extended Data Fig. 4).

The strong control of p_{O_2} on the calculated GPP values led us to investigate more restricted ranges of p_{O_2} between 0.1% and 1% PAL and between 1% and 10% PAL. In both cases the bimodality seen in the full suite of Monte Carlo calculations (Fig. 3) disappears, and both sets of calculations are well-approximated by single-peaked Gaussian probability density functions (Extended Data Fig. 5). This confirms that p_{O_2} is the dominant control on the bimodal structure seen in the full suite of Monte Carlo calculations (Fig. 3), and justifies our division of those results into a pair of Gaussian probability density functions, one associated with p_{O_2} between 0.1% and 1% PAL and another associated with p_{O_2} between 1% and 10% PAL

We fit the estimated GPP probability density to combinations of Gaussian probability density functions and found that a pair of such functions well-approximated the calculated results (Fig. 3). The parameters describing these two Gaussian probability density functions can be found in Supplementary Table 1 and indicate that our measurements of the $\Delta^{17}O$ of mid-Proterozoic sulfate are consistent with two mean values of GPP (41 ± 3% PAL and 6 ± 1% PAL, in which the uncertainty reflects 95% confidence intervals on the mean). As described above, we inverted our calculation procedure to estimate the specific p_{O_2} values that are consistent with each of these GPP estimates (Extended Data Fig. 5) and found that the high GPP state corresponds to p_{O_2} of about 2% PAL, and the low GPP state corresponds to p_{O_2} of about 0.3% PAL (Supplementary Table 1). This direct relationship between GPP and p_{O_2} is present over a wide range of GPP values (Extended Data Fig. 4b) and appears to reflect a fundamental characteristic of the mid-Proterozoic Earth.

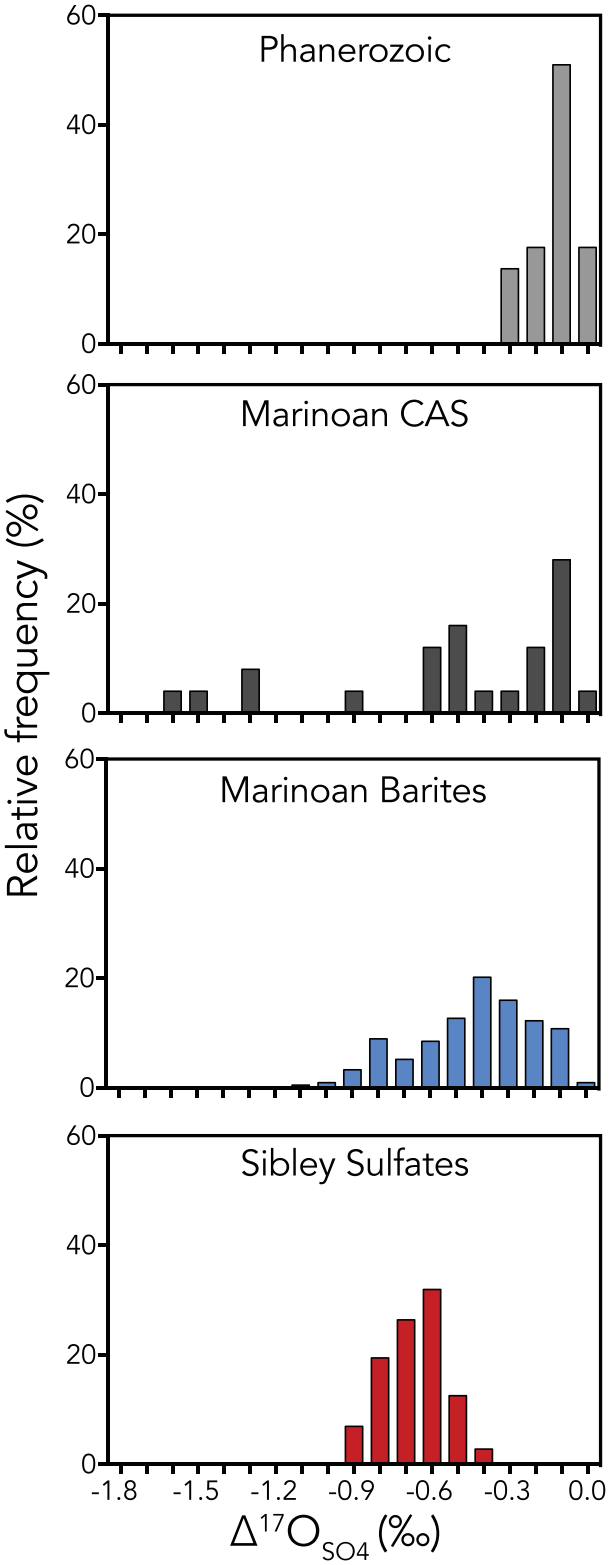
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

© 2018 Springer Nature Limited. All rights reserved.

Data availability. The authors declare that all data supporting the findings of this study are available within the paper and its Supplementary Information.

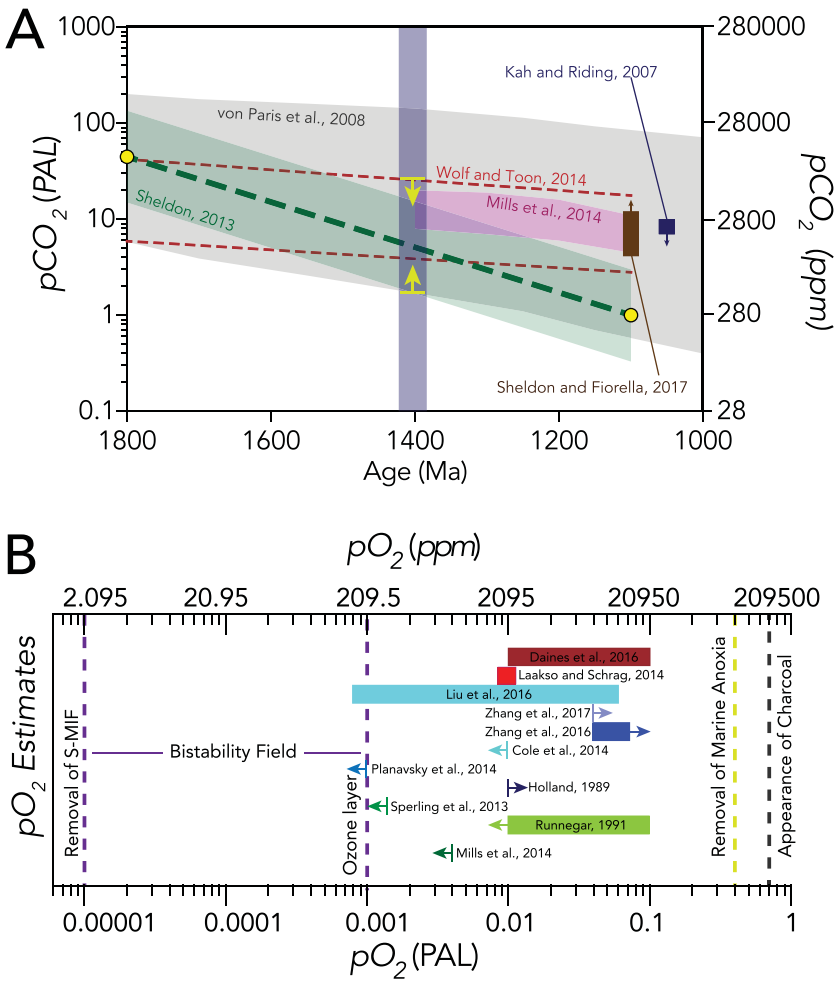
46. Davis, D. W. & Sutcliffe, R. H. U–Pb ages from the Nipigon plate and northern Lake Superior. *Geol. Soc. Am. Bull.* **96**, 1572–1579 (1985).
47. Franklin, J. M. in *Rubidium–Strontium Isotopic Age Studies, Report 2* (eds Wanless, R. K. & Loveridge, W. D.) 77–14 (Geological Survey of Canada, Ottawa, 1978).
48. Robertson, W. A. Pole position from thermally cleaned Sibley Group sediments and its relevance to Proterozoic magnetic stratigraphy. *Can. J. Earth Sci.* **10**, 180–193 (1973).
49. Elston, D. P., Enkin, R. J., Baker, J. D. & Kisilevsky, K. Tightening the belt: paleomagnetism-stratigraphic constraints on deposition, correlation, and deformation of the Middle Proterozoic (ca. 1.4 Ga) Belt-Purcell Supergroup, United States and Canada. *Geol. Soc. Am. Bull.* **114**, 619–638 (2002).
50. Cheadle, B. A. Alluvial-playa sedimentation in the lower Keweenawan Sibley Group, Thunder Bay District, Ontario. *Can. J. Earth Sci.* **23**, 527–542 (1986).
51. Metsaranta, R. T. *Sedimentology and Geochemistry of the Mesoproterozoic Pass Lake and Rossport Formations, Sibley Group*. MSc thesis, Lakehead Univ. (2006).
52. Bao, H. Purifying barite for oxygen isotope measurement by dissolution and reprecipitation in a chelating solution. *Anal. Chem.* **78**, 304–309 (2006).
53. Bao, H. & Thiemens, M. H. Generation of O_2 from $BaSO_4$ using a CO_2-laser fluorination system for simultaneous analysis of $\delta^{18}O$ and $\delta^{17}O$. *Anal. Chem.* **72**, 4029–4032 (2000).
54. Matsuhisa, Y., Goldsmith, J. R. & Clayton, R. N. Mechanisms of hydrothermal crystallization of quartz at 250 °C and 15 kbar. *Geochim. Cosmochim. Acta* **42**, 173–182 (1978).
55. Cao, X. & Liu, Y. Equilibrium mass-dependent fractionation relationships for triple oxygen isotopes. *Geochim. Cosmochim. Acta* **75**, 7435–7445 (2011).
56. Bao, H., Cao, X. & Hayles, J. A. Triple oxygen isotopes: fundamental relationships and applications. *Annu. Rev. Earth Planet. Sci.* **44**, 463–492 (2016).
57. Miller, M. F. Isotopic fractionation and the quantification of ^{17}O anomalies in the oxygen three-isotope system: an appraisal and geochemical significance. *Geochim. Cosmochim. Acta* **66**, 1881–1889 (2002).
58. Angert, A., Rachmilevitch, S., Barkan, E. & Luz, B. Effects of photorespiration, the cytochrome pathway, and the alternative pathway on the triple isotopic composition of atmospheric O_2. *Glob. Biogeochem. Cycles* **17**, 1030 (2003).
59. Thode, H. G., Monster, J. & Dunford, H. B. Sulphur isotope geochemistry. *Geochim. Cosmochim. Acta* **25**, 159–174 (1961).
60. Shaheen, R., Janssen, C. & Röckmann, T. Investigations of the photochemical isotope equilibrium between O_2, CO_2 and O_3. *Atmos. Chem. Phys.* **7**, 495–509 (2007).
61. von Paris, P. et al. Warming the early Earth—CO_2 reconsidered. *Planet. Space Sci.* **56**, 1244–1259 (2008).
62. Wolf, E. T. & Toon, O. B. Controls on the Archean climate system investigated with a global climate model. *Astrobiology* **14**, 241–253 (2014).
63. Mills, B., Lenton, T. M. & Watson, A. J. Proterozoic oxygen rise linked to shifting balance between seafloor and terrestrial weathering. *Proc. Natl Acad. Sci. USA* **111**, 9073–9078 (2014).
64. Kaufman, A. J. & Xiao, S. High CO_2 levels in the Proterozoic atmosphere estimated from analyses of individual microfossils. *Nature* **425**, 279–282 (2003).
65. Sheldon, N. D. Causes and consequences of low atmospheric pCO_2 in the Late Mesoproterozoic. *Chem. Geol.* **362**, 224–231 (2013).
66. Hansen, J. et al. Target atmospheric CO_2: where should humanity aim? *Open Atmos. Sci. J.* **2**, 217–231 (2008).
67. Kasting, J. F. & Donahue, T. M. The evolution of atmospheric ozone. *J. Geophys. Res. Oceans* **85**, 3255–3263 (1980).
68. Pavlov, A. A. & Kasting, J. F. Mass-independent fractionation of sulfur isotopes in Archean sediments: strong evidence for an anoxic Archean atmosphere. *Astrobiology* **2**, 27–41 (2002).
69. Horner, T. J. et al. Pelagic barite precipitation at micromolar ambient sulfate. *Nat. Commun.* **8**, 1342 (2017).
70. Holland, H. D., Feakes, C. R. & Zbinden, E. A. The Flin Flon paleosol and the composition of the atmosphere 1.8 BYBP. *Am. J. Sci.* **289**, 362–389 (1989).
71. Planavsky, N. J. et al. Low mid-Proterozoic atmospheric oxygen levels and the delayed rise of animals. *Science* **346**, 635–638 (2014).
72. Cole, D. B. et al. A shale-hosted Cr isotope record of low atmospheric oxygen during the Proterozoic. *Geology* **44**, 555–558 (2016).
73. Gilleaudeau, G. J. et al. Oxygenation of the mid-Proterozoic atmosphere: clues from chromium isotopes in carbonates. *Geochem. Perspect. Lett.* **2**, 178–187 (2016).
74. Zhang, S. et al. Sufficient oxygen for animal respiration 1,400 million years ago. *Proc. Natl Acad. Sci. USA* **113**, 1731–1736 (2016).
75. Planavsky, N. J. et al. No evidence for high atmospheric oxygen levels 1,400 million years ago. *Proc. Natl Acad. Sci. USA* **113**, E2550–E2551 (2016).
76. Liu, X. M. et al. Tracing Earth's O_2 evolution using Zn/Fe ratios in marine carbonates. *Geochem. Perspect. Lett.* **2**, 24–34 (2016).
77. Canfield, D. E. A new model for Proterozoic ocean chemistry. *Nature* **396**, 450–453 (1998).
78. Reinhard, C. T. et al. Proterozoic ocean redox and biogeochemical stasis. *Proc. Natl Acad. Sci. USA* **110**, 5357–5362 (2013).
79. Cox, G. M. et al. Basin redox and primary productivity within the Mesoproterozoic Roper Seaway. *Chem. Geol.* **440**, 101–114 (2016).
80. Daines, S. J., Mills, B. J. & Lenton, T. M. Atmospheric oxygen regulation at low Proterozoic levels by incomplete oxidative weathering of sedimentary organic carbon. *Nat. Commun.* **8**, 14379 (2017).
81. Runnegar, B. Precambrian oxygen levels estimated from the biochemistry and physiology of early eukaryotes. *Global Planet. Change* **5**, 97–111 (1991).
82. Sperling, E. A., Halverson, G. P., Knoll, A. H., Macdonald, F. A. & Johnston, D. T. A basin redox transect at the dawn of animal life. *Earth Planet. Sci. Lett.* **371–372**, 143–155 (2013).
83. Mills, D. B. et al. Oxygen requirements of the earliest animals. *Proc. Natl Acad. Sci. USA* **111**, 4168–4172 (2014).
84. Kohl, I. & Bao, H. Triple-oxygen-isotope determination of molecular oxygen incorporation in sulfate produced during abiotic pyrite oxidation (pH= 2–11). *Geochim. Cosmochim. Acta* **75**, 1785–1798 (2011).
85. Hayles, J. A., Cao, X. & Bao, H. The statistical mechanical basis of the triple isotope fractionation relationship. *Geochem. Perspect. Lett.* **3**, 1–11 (2017).
86. Trenberth, K. & Smith, E. L. The mass of the atmosphere: a constraint on global analyses. *J. Clim.* **18**, 864–875 (2005).
87. Linz, M. et al. The strength of the meridional overturning circulation of the stratosphere. *Nat. Geosci.* **10**, 663–667 (2017).
88. Butchart, N. et al. Simulations of anthropogenic change in the strength of the Brewer–Dobson circulation. *Clim. Dyn.* **27**, 727–741 (2006).
89. Fiorella, R. P. & Sheldon, N. D. Equable end Mesoproterozoic climate in the absence of high CO_2. *Geology* **45**, 231–234 (2017).
90. Gibson, T. M. et al. Precise age of *Bangiomorpha pubescens* dates the origin of eukaryotic photosynthesis. *Geology* **46**, 135–138 (2018).
91. Kah, L. C. & Riding, R. Mesoproterozoic carbon dioxide levels inferred from calcified cyanobacteria. *Geology* **35**, 799–802 (2007).
92. Farquhar, J., Bao, H. & Thiemens, M. Atmospheric influence of Earth's earliest sulfur cycle. *Science* **289**, 756–758 (2000).
93. Goldblatt, C., Lenton, T. M. & Watson, A. J. Bistability of atmospheric oxygen and the Great Oxidation. *Nature* **443**, 683–686 (2006).
94. Zhang, S. et al. The oxic degradation of sedimentary organic matter 1400 Ma constrains atmospheric oxygen levels. *Biogeosciences* **14**, 2133–2149 (2017).

© 2018 Springer Nature Limited. All rights reserved.



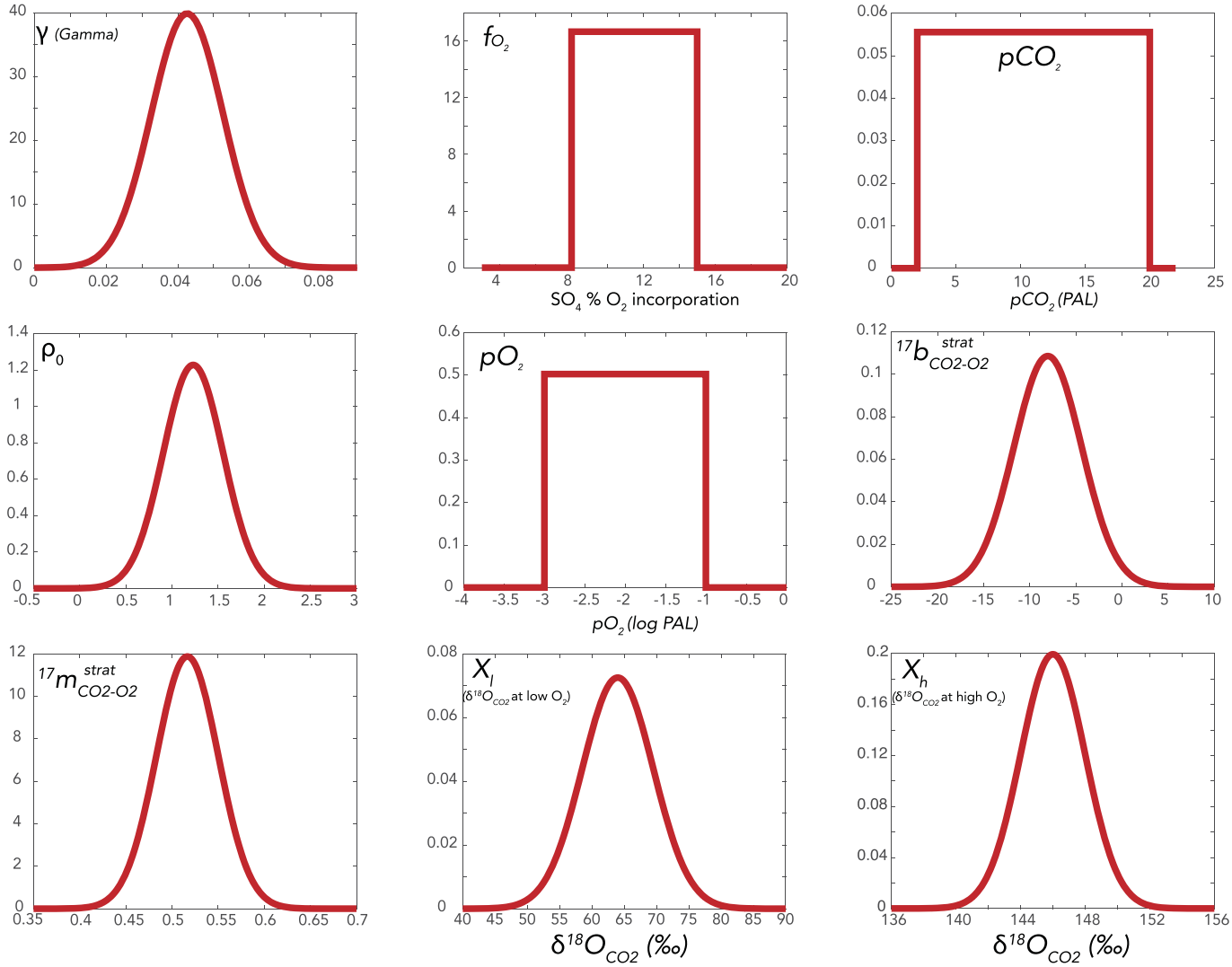
Extended Data Fig. 1 | Histograms of existing data for $\Delta^{17}O$ values through Earth history. The histograms show Phanerozoic sulfates[2] ($n = 51$; light grey), syn-Marinoan CAS[29] ($n = 25$; dark grey), post-Marinoan barite[2,39–42] ($n = 213$; blue) and results from the Sibley sulfates ($n = 68$; red).

© 2018 Springer Nature Limited. All rights reserved.



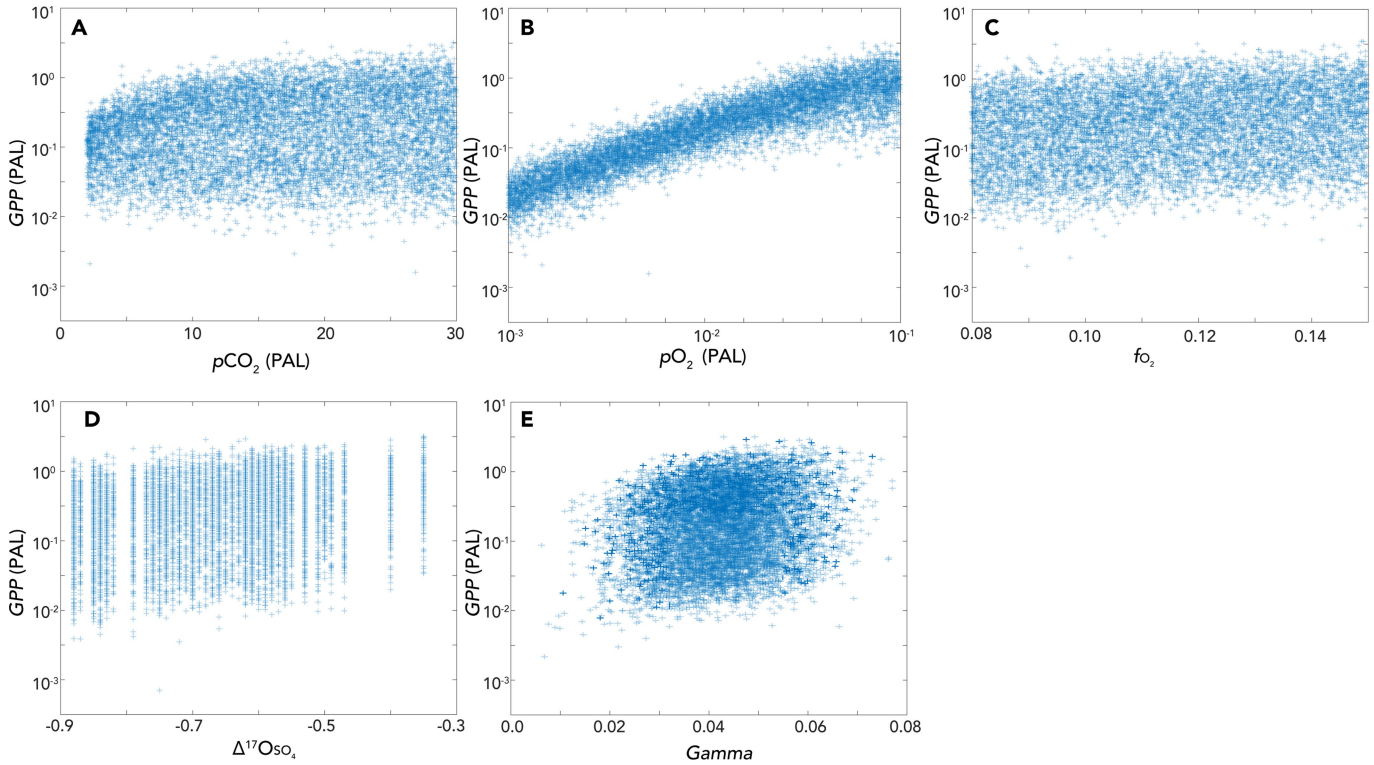
Extended Data Fig. 2 | Compiled p_{CO_2} and p_{O_2} estimates. a, p_{CO_2} estimates. Left *y* axis, percentage of PAL; right *y* axis, p.p.m. Grey band outlines results from 1D modelling[61] based on temperatures of 273 K (bottom), 288 K (top) and changing solar luminosity. Red dotted lines represent extrapolated general circulation model (GCM) results[62] from Archaean estimates. The green-shaded region represents the uncertainty envelope of palaeosol-based estimates[65] with the green dotted lines interpolating between estimates at 1.8 and 1.1 Ga together. The pink-shaded region represents estimates based on the COPSE Earth system model[63]. The brown bar represents modelling-based estimates required to prevent a global glaciation at 1.1 Ga[89]. The dark-blue square is the microfossil-based estimate that sets maximum limits at 1.05 Ga[90,91]. Yellow arrows represent the upper (30 PAL) and lower (2 PAL) limits used in this work. Data are from previous publications[61–63,65,89,91]. **b**, p_{O_2} estimates. Green arrows represent biologically based estimates; blue arrows represent geochemical estimates; and in red are modelling p_{O_2} estimates. Purple lines represent the removal of S-MIF[92], a proposed bistability field[93], and constraints on the establishment of a modern-like ozone layer[13,67]. The yellow dashed line represents the suggested limits for the removal of deep ocean anoxia[76], and the grey dashed line represents the appearance of charcoal. Data are from previous publications[9,19,67,68,70–72,74,76,77,80–83,93,94].

© 2018 Springer Nature Limited. All rights reserved.



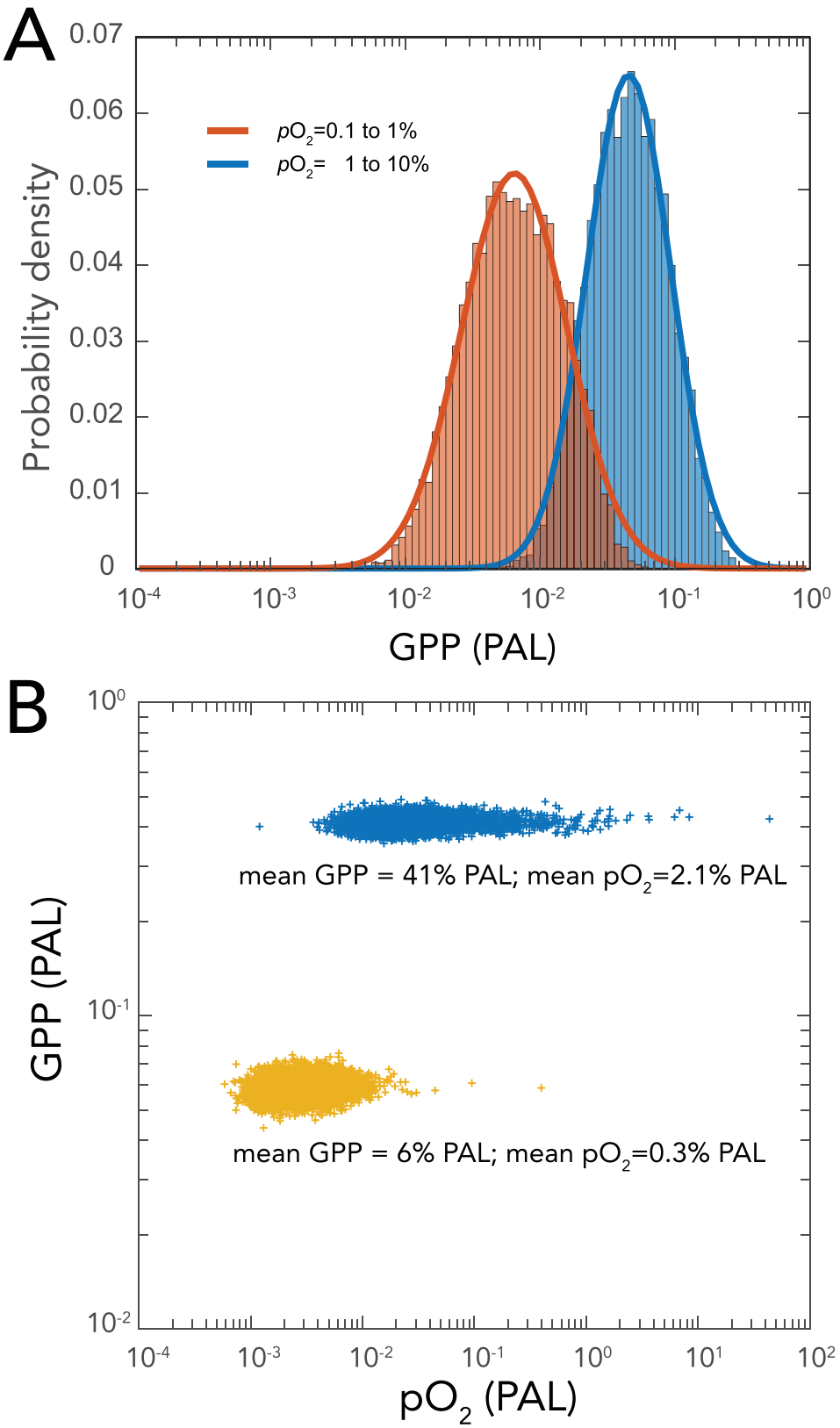
Extended Data Fig. 3 | Probability distribution functions for the control parameters of the model. The justification for the form (Gaussian or uniform), the spread (standard deviation or range) and the mean of each distribution is given in the main text and Methods.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 4 | Realization of about 10,000 Monte Carlo calculations of GPP (PAL) relative to various control parameters in the isotope mass-balance model. a, Unlike model calculations for the Neoproterozoic era, there is no clear strong dependence of GPP on assumed p_{CO_2}. **b**, There is a clear log-linear dependence of the GPP estimates on p_{O_2}. **c**, GPP responds weakly to f_{O_2}, with large fractions of O_2 in sulfate (which indicate smaller $\Delta^{17}O_{O2}$ values) leading to higher estimates of GPP. **d**, The response of GPP to $\Delta^{17}O_{sulfate}$ is similar to the response to f_{O_2}. Smaller $\Delta^{17}O_{sulfate}$ values indicate smaller $\Delta^{17}O_{O2}$ values, which—all else being the same—requires greater GPP. **e**, GPP estimates seem to be largely independent of gamma.

© 2018 Springer Nature Limited. All rights reserved.

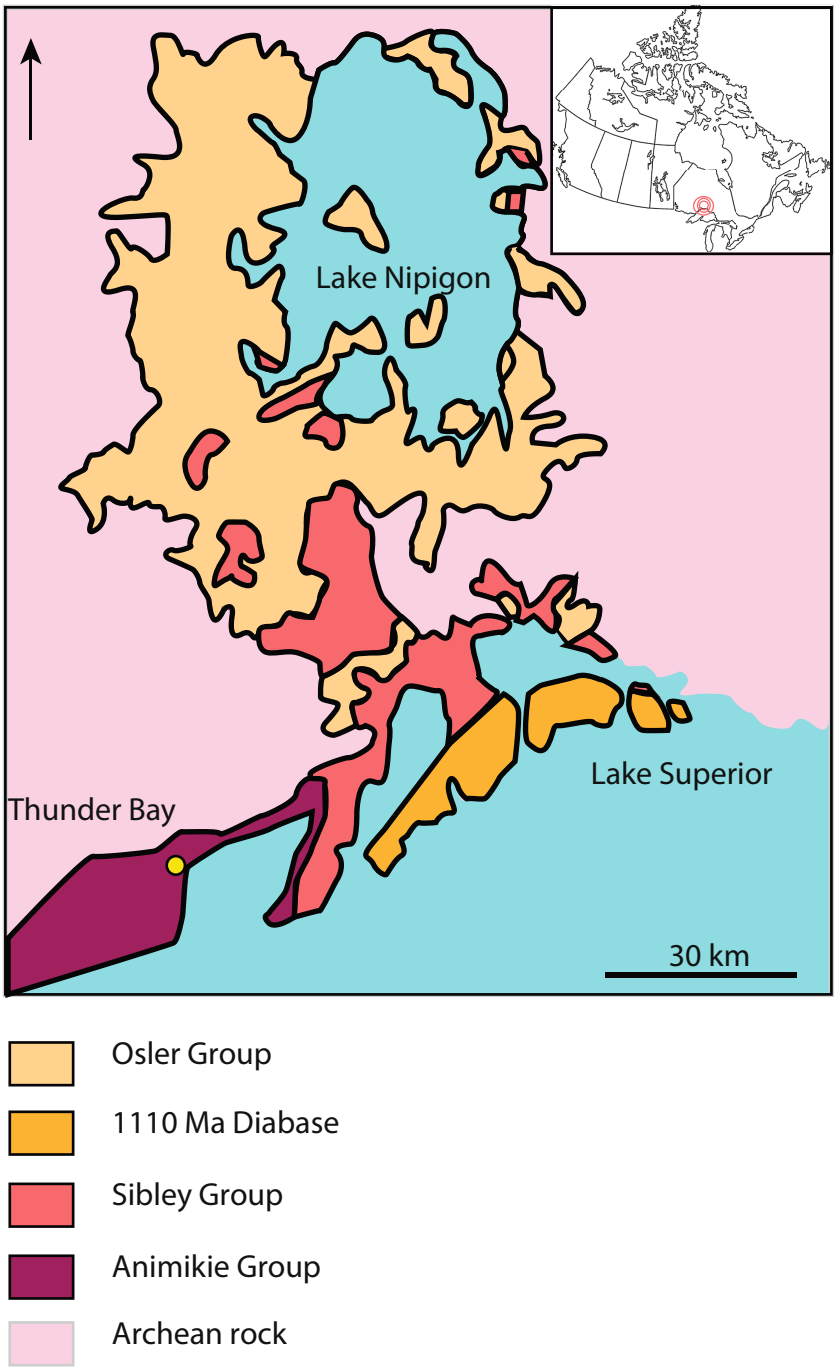


Extended Data Fig. 5 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

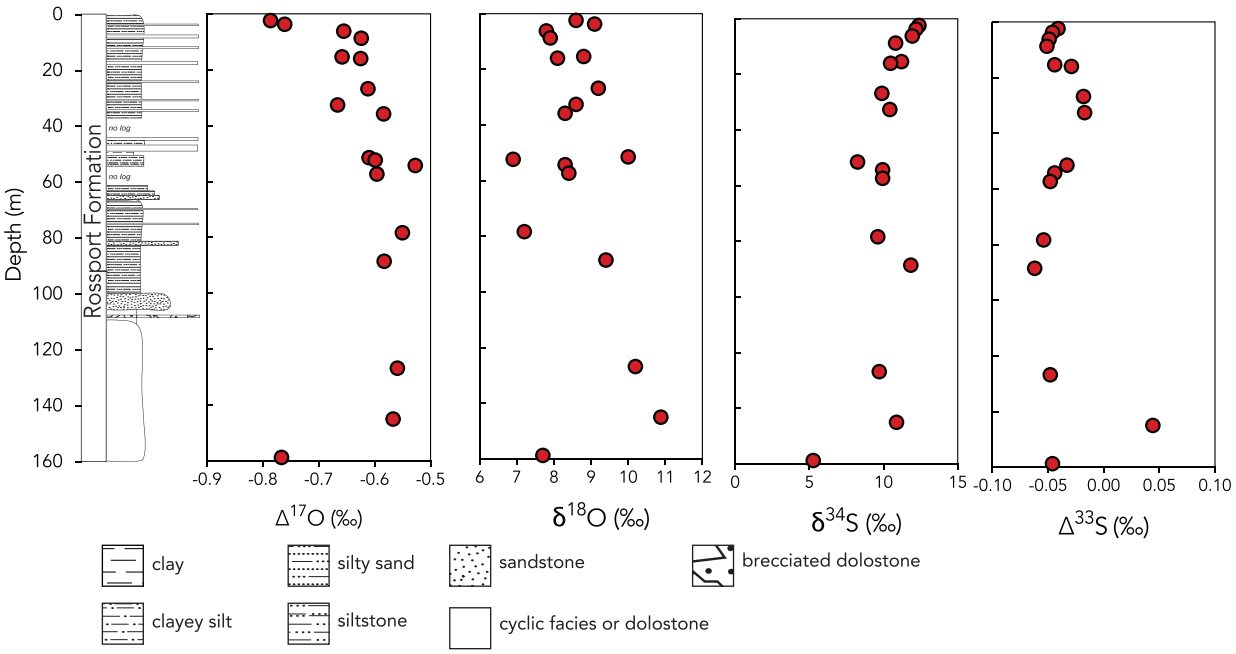
Extended Data Fig. 5 | Realization of about 20,000 Monte Carlo calculations of GPP (PAL) and p_{O_2}. a, Results are calculated to be consistent with the $\Delta^{17}O$ measurements in Extended Data Table 1 and the probability distribution functions shown in Extended Data Fig. 2, with the exception of p_{O_2}. These calculations assume restricted ranges of p_{O_2} between 0.1% and 1% PAL (orange histogram) and between 1% and 10% PAL (blue histogram). In both cases the bimodality seen in the full suite of Monte Carlo calculations (Fig. 3) disappears, and both sets of calculations are well-approximated by single-peaked Gaussian probability density functions. This confirms that p_{O_2} is the dominant control on the bimodal structure seen in the full suite of Monte Carlo calculations (Fig. 3), and justifies our division of those results into a pair of Gaussian probability density functions, one associated with p_{O_2} between 0.1% and 1% PAL and another associated with p_{O_2} between 1% and 10% PAL. **b**, Results are solutions that are consistent with the $\Delta^{17}O_{sulfate}$ dataset (Extended Data Table 1), the probability density functions shown in Extended Data Fig. 1 (with the exception of p_{O_2}) and the mean GPP estimates of the two inferred Gaussian probability distributions for mid-Proterozoic GPP (Fig. 3). GPP was allowed to vary in a Gaussian fashion between the 95% confidence limits on the GPP mean values.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 6 | Geological map of the Lake Nipigon–northern Lake Superior region. This figure was adapted from a previous publication[28].
© 2008 Canadian Science Publishing or its licensors. Reproduced with permission.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 7 | Oxygen and sulfur isotope compositions ($\Delta^{17}O$, $\delta^{18}O$, $\delta^{34}S$ and $\Delta^{33}S$) for sulfates from drill hole NI-92-7 plotted against stratigraphic height. Uncertainty on all analyses is smaller than the sizes of the data points.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Isotopic data and comparisons with previously published results

A

Sample	$\Delta^{17}O$	$\delta^{18}O$	$\delta^{34}S$	$\Delta^{33}S$	Sample	$\Delta^{17}O$	$\delta^{18}O$	$\delta^{34}S$	$\Delta^{33}S$
PF-5	-0.62	6.9	6.2	-0.027	NI-92-7-15.4	-0.66	8.8	11.2	-0.044
PF-12	-0.67	8.8	10.6	-0.039	NI-92-7-15.95	-0.62	8.1	10.5	-0.029
PF-2	-0.53	9.8	12.3	-0.040	NI-92-7-26.8	-0.61	9.2	9.9	-0.018
PF-20	-0.56	1.0	9.4	-0.038	NI-92-7-32.58	-0.66	8.6	10.4	-0.017
PF-11	-0.63	8.3	12.4	-0.033	NI-92-7-35.75	-0.58	8.3		
PF-21	-0.59	10.9	10.6	-0.042	NI-92-7-51.45	-0.61	10.0	8.3	-0.033
PF-4	-0.61	7.9	11.6	-0.043	NI-92-7-52.3	-0.60	6.9		
PF-10	-0.50	10.9	11.4	-0.047	NI-92-7-54.25	-0.53	8.3	9.9	-0.044
PF-7	-0.53	8.3	11.6	-0.050	NI-92-7-57.3	-0.60	8.4	9.9	-0.048
PF-8	-0.47	8.4	10.1	-0.054	NI-92-7-78.3	-0.55	7.2	9.6	-0.054
PF-9	-0.88	6.9	8.9	-0.027	NI-92-7-88.5	-0.58	9.4	11.8	-0.062
PF-9 re-run	-0.70				NI-92-7-126.75	-0.56	10.2	9.7	-0.048
PF-6	-0.68	6.0	9.1	-0.036	NI-92-7-158.7	-0.77	7.7	5.3	-0.046
PF-6 re-run	-0.66				NI-92-7-164.7				
PF-15	-0.58	9.7	12.0	-0.053	WP-07-03-1	-0.69	6.5		
PF-13	-0.74	10.8	12.0	-0.037	WP-07-03-2	-0.84	6.9	5.6	-0.045
PF-13 re-run	-0.75				WP-07-03-3	-0.58	9.1		
PF-1	-0.59	11.1	12.0	-0.043	WP-07-03-4	-0.49	8.0		
PF17	-0.40	7.6	13.0	-0.039	WP-07-03-5	-0.79	7.3		
PF-3	-0.64	13.1	12.1	-0.048	WP-07-03-8	-0.70	6.6		
PF-3 re-run	-0.56				WP-07-03-10	-0.76	8.5		
PF-16	-0.35	9.6	13.5	-0.039	WP-07-03-10	-0.88	6.8	5.9	-0.220
PF-18	-0.67	8.3	9.8	-0.034	WP-07-03-11	-0.71	6.8		
PF-19	-0.57	10.9	10.7	0.044	WP-07-03-12	-0.82	6.6		
PF-18-	-0.59	13.3	11.9	-0.052	WP-07-03-12	-0.72	6.6		
03RM26	-0.51	12.9	9.3	-0.037	WP-07-03-13	-0.83	7.0		
03RM27	-0.75	3.3	10.6	-0.044	WP-07-03-15	-0.85	7.0		
03Rm62	-0.85	4.0	11.4	-0.039	WP-07-03-18	-0.84	8.1		
04Rm9	-0.84	2.9	8.8	-0.039	WP-07-03-19	-0.87	8.4	4.7	-0.41
04RM30	-0.74	5.9	5.3	-0.044	WP-07-03-20	-0.73	8.4		
04RM31	-0.72	9.8	5.4	-0.032	WP-07-03-21	-0.85	8.9	6.4	-0.24
NI-92-7-2.35	-0.79	8.6	12.4	-0.041	WP-07-03-23	-0.70	8.8		
NI-92-7-3.7	-0.76	9.1	12.2	-0.046	WP-07-03-216.6	-0.88	7.7		
NI-92-7-6.2	-0.65	7.8	12.0	-0.049	WP-07-03-217.6	-0.83	6.5		
NI-92-7-8.72	-0.62	7.9	10.8	-0.051					

B

	This Study				Marinoan Barite	Marinoan CAS
	$\Delta^{17}O$-$\delta^{18}O$	$\delta^{18}O$-$\delta^{34}S$	$\Delta^{17}O$-$\delta^{34}S$	$\Delta^{33}S$-$\delta^{34}S$	$\Delta^{17}O$-$\delta^{34}S$	$\Delta^{17}O$-$\delta^{34}S$
F	9.43	3.64	16.22	1.16	45.2	238.1
P value	0.003	0.063	0.0002	0.29	0.0003	<0.0001
R^2	0.13	0.076	0.27	0.025	0.87	0.96
Significant?	Yes	No	Yes	No	Yes	Yes

C

Samples	This Study	Phanerozoic Evaporites	Cryogenian Barites	Cryogenian CAS
Total number of values	68	51	213	25
Minimum	-0.88	-0.34	-1.05	-1.64
25% Percentile	-0.76	-0.18	-0.57	-0.63
Median	-0.66	-0.12	-0.42	-0.36
75% Percentile	-0.58	-0.06	-0.26	-0.12
Maximum	-0.35	0.00	-0.02	-0.04
Mean	-0.67	-0.13	-0.44	-0.51
Std. Dev.	0.12	0.08	0.23	0.48
Std. Err.Mean	0.01	0.01	0.02	0.10
-95% CI Mean	-0.71	-0.16	-0.47	-0.71
+95% CI Mean	-0.64	-0.11	-0.40	-0.31

D

Sample	$\Delta^{17}O$	Description
Dresser Mine	-0.076	≈3.5 Ga barite samples from North Pole Dome, West Australia[69]
PCNP-1	-0.039	
PCNP-2	-0.069	
PCNP-4	0.083	
Mean	-0.025	95% Error on Mean = 0.080

a, Data for Sibley samples measured in this study. Associated errors on total laboratory procedures and analysis are presented in the Methods. **b**, Results of a linear regression analysis for $\Delta^{17}O$–$\delta^{18}O$ ($n = 64$), $\delta^{18}O$–$\delta^{34}S$ ($n = 46$), $\Delta^{17}O$–$\delta^{34}S$ ($n = 46$) and $\Delta^{33}S$–$\delta^{34}S$ ($n = 46$) data from this study, along with results from previously published syn-Marinoan CAS[29], and post-Marinoan barite[2,39–42]. Correlations (Pearson's) are significant if $P < 0.05$. **c**, Summary statistics of $\Delta^{17}O$ results on samples from this study compared to post-Marinan barites[2,39–42], syn-Marinoan CAS[29] and Phanerozoic evaporites[2]. Samples were binned in 0.1‰ increments. **d**, New $\Delta^{17}O$ measurements from Archaean barites. Data are from a previous publication[69].

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 2 | Summary of reference parameters for model calculations

Parameter	Definition	Value	Ref.
$pO_{2,\,ref}$	pO_2 at 1 Pre-Anthropogenic Level (PAL)	1 PAL = 209500 ppmv	
$pCO_{2,\,ref}$	pCO_2 at 1 Pre-Anthropogenic Level (PAL)	1 PAL = 280 ppmv	
ρ_{ref}	Reference pO_2 -pCO_2 ratio at 1 Pre-Anthropogenic Level (PAL)	748.21	see text
τ_{ref}	Pre-anthropogenic O_2 residence time	1244 yr	18
$\Delta^{17}O_{modern\ O_2}$	$\Delta^{17}O$ of modern O_2 using arithmetic definition for $\Delta^{17}O$ and slope of 0.5305	-0.546‰	25
f_{O_2}	Fraction of O_2 incorporation during sulfide oxidation	0.08-0.15	22
γ	Rate constant for stratosphere-troposphere exchange	0.0426 yr^{-1}	see text
θ	Fraction of O_2 in stratospheric O_2-O_3-CO_2 cycles	0.1156	see text
ρ_0	pO_2 -pCO_2 ratio in O_2-O_3-CO_2 isotope fractionation experiments	1.23	60
X_l	$\delta^{18}O_{CO_2\text{-}O_2}$ value in O_2-O_3-CO_2 isotope fractionation experiments at low pO_2	64‰	60
X_h	$\delta^{18}O_{CO_2\text{-}O_2}$ value in O_2-O_3-CO_2 isotope fractionation experiments at high pO_2	146‰	60
$^{17}m^{strat}_{CO_2\text{-}O_2}$	Slope of fit to $\delta^{18}O_{CO_2\text{-}O_2}$ and $\Delta^{17}O_{CO_2\text{-}O_2}$ using arithmetic definition for $\Delta^{17}O$ and slope of 0.5305	0.5167	3,60, see text
$^{17}b^{strat}_{CO_2\text{-}O_2}$	Intercept of fit to $\delta^{18}O_{CO_2\text{-}O_2}$ and $\Delta^{17}O_{CO_2\text{-}O_2}$ using arithmetic definition for $\Delta^{17}O$ and slope of 0.5305	-8.025‰	3,60, see text

Data are from previous publications[3,18,22,25,60].

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0344-3

The outer membrane is an essential load-bearing element in Gram-negative bacteria

Enrique R. Rojas[1,2,3], Gabriel Billings[4], Pascal D. Odermatt[1,5], George K. Auer[6], Lillian Zhu[1], Amanda Miguel[1], Fred Chang[5], Douglas B. Weibel[6,7,8], Julie A. Theriot[2,3,9,10] & Kerwyn Casey Huang[1,3,10,11]*

Gram-negative bacteria possess a complex cell envelope that consists of a plasma membrane, a peptidoglycan cell wall and an outer membrane. The envelope is a selective chemical barrier[1] that defines cell shape[2] and allows the cell to sustain large mechanical loads such as turgor pressure[3]. It is widely believed that the covalently cross-linked cell wall underpins the mechanical properties of the envelope[4,5]. Here we show that the stiffness and strength of *Escherichia coli* cells are largely due to the outer membrane. Compromising the outer membrane, either chemically or genetically, greatly increased deformation of the cell envelope in response to stretching, bending and indentation forces, and induced increased levels of cell lysis upon mechanical perturbation and during L-form proliferation. Both lipopolysaccharides and proteins contributed to the stiffness of the outer membrane. These findings overturn the prevailing dogma that the cell wall is the dominant mechanical element within Gram-negative bacteria, instead demonstrating that the outer membrane can be stiffer than the cell wall, and that mechanical loads are often balanced between these structures.

The three essential layers of the Gram-negative cell envelope (Fig. 1a) are chemically and structurally diverse: the plasma membrane is a fluid phospholipid bilayer, the peptidoglycan cell wall is a covalently cross-linked macromolecule, and the outer membrane possesses phospholipids in the inner leaflet and lipopolysaccharides (LPS) in the outer leaflet. A primary role of the envelope is to sustain mechanical forces[3], and it is universally assumed that the mechanical integrity of the envelope is conferred by the cell wall[4,5]. However, the unique chemistry of the outer membrane leads to notable physical properties. For example, although proteins freely diffuse in the plasma membrane, the motion of outer-membrane proteins is constrained[6–8]. In this light, we investigated whether the outer membrane contributed to the mechanics of the cell envelope.

To assay the mechanical properties of the *E. coli* envelope, we first measured its contraction when turgor pressure (approximately 1 atm[3,9]) was eliminated by subjecting cells to a large hyperosmotic shock[10]. This shock induced plasmolysis[11] whereby the inner membrane receded from the cell wall (Fig. 1b, Supplementary Video 1), indicating that the complex formed by the cell wall and the outer membrane had contracted to its relaxed state (Extended Data Fig. 1). Plasmolysis caused the length of the cell wall to contract by $\varepsilon_p = 9.6 \pm 2.9\%$ ($\varepsilon_p = (l_1 - l_2)/l_2$, in which l_1 is the pre-shock length and l_2 is the post-shock length; mean ± s.d.; Fig. 1c, d). To test whether the outer membrane affected this response, we damaged it by exposing plasmolysed cells to a detergent (5% *N*-lauroyl sarcosine; Fig. 1b, Extended Data Fig. 2a–d, Supplementary Videos 1, 2). Cells lysed within minutes of detergent treatment (Fig. 1b, Supplementary Video 1), coinciding with a brief phase of rapid cell-wall expansion (Fig. 1c) that was probably caused by viscous drag on the cell wall by the cytoplasm as it was released from the cells (Supplementary Information, Extended Data Fig. 2e, f). Cell-wall length then contracted beyond its post-plasmolysis value by $\varepsilon_l = 14.5 \pm 8.3\%$ ($\varepsilon_l = (l_2 - l_{cw})/l_{cw}$, in which l_{cw} is the cell-wall length after detergent treatment; Fig. 1e, Supplementary Video 1). Thus, although plasmolysis caused the cell wall to contract, compromising the outer membrane caused a second, larger, contraction ($P = 10^{-6}$, Student's two-sided *t*-test) that was not due to changes in cell-wall composition (Extended Data Fig. 2g) or compaction by the detergent (Supplementary Information, Extended Data Fig. 3a, b). We observed similar behaviour under other growth conditions, in other *E. coli* strains and in other Gram-negative species, but not in the Gram-positive bacterium *Bacillus subtilis* (Supplementary Information, Extended Data Fig. 3b–f).

Under turgid conditions, the cell wall is under extreme extension: between the turgid state and the fully relaxed state, the cell-wall length contracted by a total of $\varepsilon_t = 25.0 \pm 8.6\%$ ($\varepsilon_t = (l_1 - l_{cw})/l_{cw}$; $n = 56$), with increased contraction at higher detergent concentrations (Extended Data Fig. 4a). In addition, total contraction was correlated with the residual phase density of the cell after lysis (Fig. 1b, arrow), which was caused by retention of specific proteins within the sacculus (Extended Data Fig. 4b–h, Supplementary Information). The least phase-dense cells contracted by as much as 50% (Fig. 1f, Extended Data Fig. 4h). These data suggest that after lysis, residual cytoplasm within the envelope caused an entropic, turgor-like pressure within cells, indicating that our measurement of the contraction upon lysis was actually lower than it would have been if all cytoplasmic contents were lost. By comparison to the cell wall, the relative length extension at which typical materials plastically deform ranges from approximately 0.01% to approximately 5% for pure elements[12], and is approximately 10% for agarose gels[13].

Our results suggested that the outer membrane was stabilizing the cell wall in a highly stretched state during plasmolysis by bearing compressive stress, thereby balancing tensile stress in the wall (Fig. 1c, right). This model implies that the relaxed size of the outer membrane is larger than that of the cell wall (and larger than the size of the cell envelope after plasmolysis) and that the outer membrane can bear mechanical forces comparable to those borne by the wall. To estimate the rest length of the outer membrane, we plasmolysed cells and then digested their cell walls with lysozyme, thereby allowing their outer membranes to relax (Fig. 2a). As cell-wall digestion caused the cells to form spheroplasts (Fig. 2a), we measured the surface area of the outer membrane and calculated the length that it would have had in a rod-like shape, given this surface area. We found that the rest length of the outer membrane (l_{om}) was precisely equal to the length of the turgid cell ($l_{om}/l_1 = 1.0 \pm 0.11$ for *E. coli* MG1655, $n = 12$, Fig. 2b, with similar values for other wild-type strains, Extended Data Fig. 3g), and was therefore larger than the length after plasmolysis. Thus, during steady-state growth, the outer membrane bears little or no load. Using this

[1]Department of Bioengineering, Stanford University, Stanford, CA, USA. [2]Department of Biochemistry, Stanford University School of Medicine, Stanford, CA, USA. [3]Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA, USA. [4]Department of Physics, Stanford University, Stanford, CA, USA. [5]Department of Cell and Tissue Biology, University of California San Francisco, San Francisco, CA, USA. [6]Department of Biomedical Engineering, University of Wisconsin-Madison, Madison, WI, USA. [7]Department of Biochemistry, University of Wisconsin-Madison, Madison, WI, USA. [8]Department of Chemistry, University of Wisconsin-Madison, Madison, WI, USA. [9]Howard Hughes Medical Institute, Stanford, CA, USA. [10]Biophysics Program, Stanford University, Stanford, CA, USA. [11]Chan Zuckerberg Biohub, San Francisco, CA, USA. *e-mail: kchuang@stanford.edu

© 2018 Springer Nature Limited. All rights reserved.

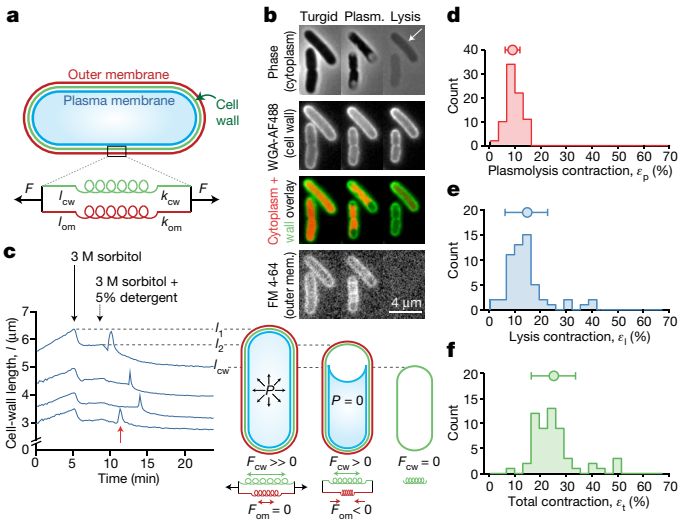


Fig. 1 | Detergent treatment after plasmolysis causes further contraction of the Gram-negative cell wall. a, Model of the cell wall–outer membrane complex as parallel linear springs with spring constants k_{cw}, k_{om}, and rest lengths l_{cw}, l_{om}. **b**, *E. coli* cells (turgid, plasmolysed (Plasm.), lysed) stained with WGA-488 and FM 4-64. The white arrow indicates residual phase signal after lysis. Images are representative of 84 cells from 3 experiments. Outer mem., outer membrane. **c**, Left, cell-wall length versus time during hyperosmotic shock and treatment with detergent. Data are representative of 84 cells. The red arrow indicates sharp swelling upon lysis. Right, a model of the turgid, plasmolysed and lysed cellular states. **d**–**f**, Histograms of length contraction upon plasmolysis (**d**, $n = 79$ cells), lysis (**e**, $n = 56$ cells) and in total (**f**, $n = 56$ cells). Circle and error bars indicate the mean ± s.d.

finding, we solved for the stiffness of the outer membrane by treating the cell wall–outer membrane complex as parallel linear springs

$$k_{om} = \frac{\varepsilon_l}{\varepsilon_p(\varepsilon_l + 1)} k_{cw} \quad (1)$$

from which it follows that the outer membrane is stiffer than the cell wall, with $k_{om} = 1.32k_{cw}$.

We thus sought to dissect the molecular basis underlying the load-bearing capacity of the outer membrane by altering its composition via chemical and genetic means. Divalent cations mediate ionic bonds between negatively charged moieties within LPS, including charges in the lipid A and the core oligosaccharide domains[14]. Notably, we found that *Vibrio cholerae* O1, which has one phosphate group within lipid A substituted[15], has a relatively compliant outer membrane (Extended Data Fig. 3f). Accordingly, we hypothesized that ionic bonds were important for the load-bearing capacity of the outer membrane and that disrupting these bonds would decrease its stiffness. Treating cells with EDTA, which chelates magnesium, released LPS from the outer membrane[16,17] (Extended Data Fig. 2d), and had no direct effect on cell-wall length (Extended Data Fig. 5a–c) or composition (Extended Data Fig. 2g). We therefore predicted that EDTA treatment after plasmolysis would cause a cell-wall contraction similar to detergent treatment. Indeed, the walls of plasmolysed cells began to contract shortly after EDTA was added (Fig. 2c) and the degree of contraction increased with EDTA concentration (Extended Data Fig. 5d). Adding saturating magnesium blocked this effect (Extended Data Fig. 5e). When we repeated the plasmolysis and lysis experimental sequence (Fig. 1c) with the addition of 10 mM EDTA in both the plasmolysis and lysis media, cell walls contracted: i) much more upon plasmolysis (Fig. 2d); ii) much less upon detergent treatment (Fig. 2e) and iii) the same total amount as untreated cells (Fig. 2f), reducing the estimated stiffness of the outer membrane by 70% (Fig. 2g).

The dye FM 4-64, which intercalates into the outer membrane, did not have a significant effect on outer-membrane stiffness (Fig. 2d–g), whereas genetic alterations to the outer membrane had marked effects. A mutant *E. coli* strain with the *imp4213* allele of *lptD*, encoding a component of the LPS assembly machinery, has a porous outer membrane rich in phospholipids[18,19]. As with EDTA-treated cells, *imp4213* cells contracted more upon plasmolysis (Fig. 2d) and less upon lysis (Fig. 2e), with no change in total contraction (Fig. 2f), indicating reduced outer-membrane stiffness (Fig. 2g). A mutant lacking the abundant outer-membrane protein OmpA exhibited a similar phenotype (Fig. 2d–g), demonstrating that proteins also contribute to outer-membrane mechanics.

The LPS of most Gram-negative bacteria include a polysaccharide component called the O antigen (although many domesticated strains do not). Notably, adding the O8 antigen to a domesticated *E. coli* strain (AB1133) that exhibited no contraction upon lysis (Extended Data Fig. 3h) greatly reduced its contraction upon plasmolysis (Fig. 2d, Extended Data Fig. 3h) and increased its contraction upon lysis (Fig. 2e, Extended Data Fig. 3h), suggesting that the antigen greatly stiffens the outer membrane (Fig. 2g). The O8 antigen is an electrically neutral poly-mannose molecule, demonstrating that non-ionic interactions between LPS molecules can also contribute to outer-membrane stiffness.

For the outer membrane to exert force on the cell wall during plasmolysis, there must be mechanical coupling between these structures. Deletion of *pal*—a gene that encodes a component of the Tol–Pal complex, which bridges the outer membrane and the cell wall—resulted in cell walls that underwent extreme contraction upon plasmolysis (Fig. 2d) and no contraction upon lysis (Fig. 2e), implicating this complex functions as a mechanical linker. Braun's lipoprotein (Lpp) is a highly abundant protein that also binds both the cell wall and the outer membrane. Mutant cells lacking *lpp* contracted more upon both plasmolysis and lysis, yielding a very large total contraction (Fig. 2d–f). Because the estimate for the ratio between the stiffness of the cell wall and the outer membrane did not change (Fig. 2g), but the total contraction increased markedly (Fig. 2f), we speculate that deletion of *lpp* affects the coupling between the cell wall and the outer membrane as well as the stiffness of the cell wall and/or the outer membrane directly.

To confirm that our measurements of envelope stiffness were not affected by variation in initial turgor pressure between conditions or strains, we subjected *E. coli* cells to small oscillatory hyperosmotic shocks (Fig. 2h, Supplementary Video 3), which caused cells and their envelopes to oscillate in size as they grew (Fig. 2i) but caused negligible plasmolysis[10]. We hypothesized that the amplitude of cell-envelope oscillations should depend on envelope stiffness and should therefore be affected by perturbations to the outer membrane. In the absence of outer-membrane perturbation, cell-wall length oscillated with an amplitude of 2.2 ± 0.2%. Treating cells with EDTA during the oscillations starkly increased the amplitude of the oscillations (Fig. 2i, j, Extended Data Fig. 5f–k), as did FM 4-64 (Fig. 2j). As predicted, mutations that decreased outer-membrane stiffness (*imp4213* and Δ*ompA*) increased the oscillation amplitude (Fig. 2j, Extended Data Fig. 5l), whereas the O8 antigen decreased the amplitude (Fig. 2j). Deletion of *lpp* increased the oscillation amplitude, whereas deletion of *pal* did not (Fig. 2j). These data support our model that although Lpp weakens the outer membrane and/or the cell wall as well as their coupling, the Tol–Pal complex is responsible only for mechanical coupling, which is less important during small shocks that do not plasmolyse the cells[10].

Each of the measurements described above tested the mechanical properties of the outer membrane under compression. *E. coli* cells do not swell under hypoosmotic shock[10], limiting our ability to test whether the outer membrane can bear tensile loads (although a very stiff outer membrane could account for this observation). To determine whether the outer membrane affects the mechanical properties of the cell envelope under other types of load, we used atomic force microscopy (AFM) to directly measure the stiffness of turgid cells, which is dependent on envelope stiffness[3]. Progressive addition of EDTA or FM 4-64 decreased cell stiffness in a concentration-dependent manner (Fig. 3a, b, Extended Data Fig. 6a, b). Mutations

© 2018 Springer Nature Limited. All rights reserved.

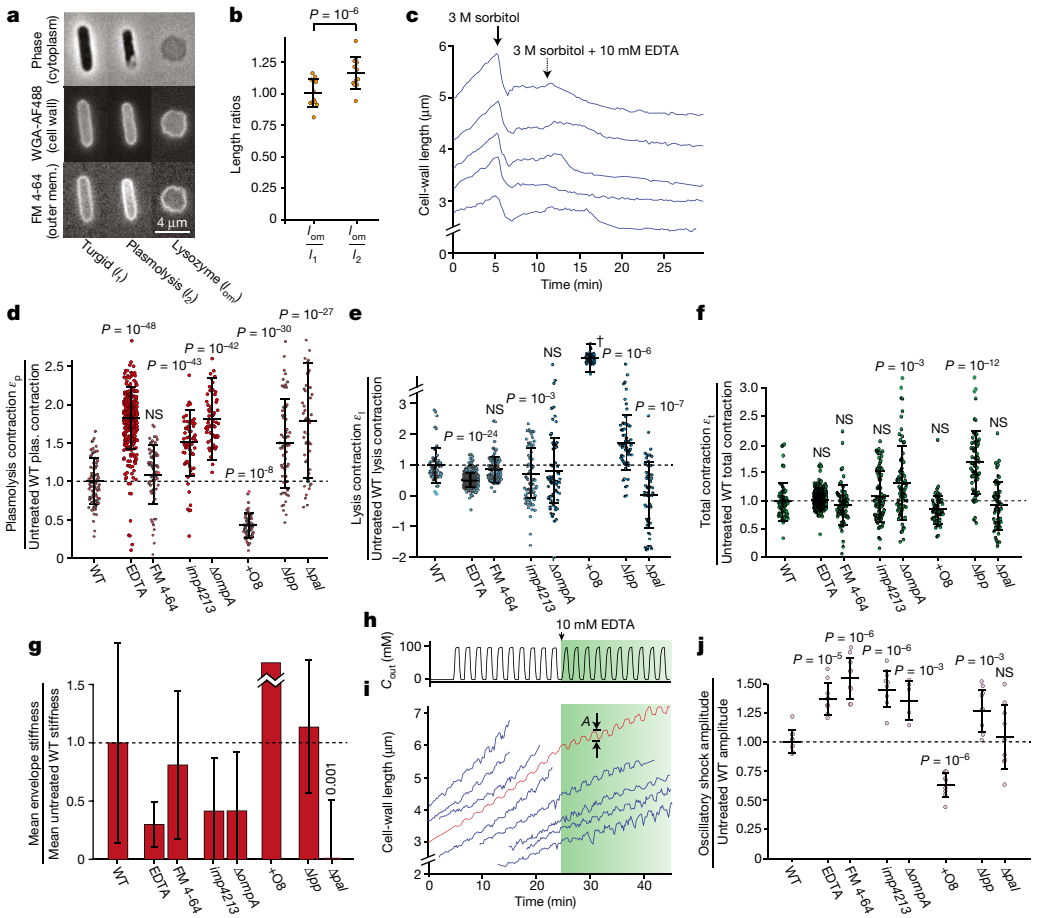


Fig. 2 | Cellular mechanical properties are dependent on the composition and integrity of the outer membrane. a, *E. coli* cells (turgid, plasmolysed and lysozyme-treated) stained as in Fig. 1b (one experiment, similar results for two other wild-type strains, Extended Data Fig. 3g). **b**, Mean ratios between the rest length of the outer membrane, l_{om}, and the length of the turgid cell, l_1, and between l_{om} and the length of the cell envelope during plasmolysis, l_2. Data are mean ± s.d. from 12 cells. *P* value from a paired two-sided Student's *t*-test. **c**, Length of representative cell walls versus time during hyperosmotic shock and treatment with EDTA ($n = 184$ cells total). **d**–**f**, Cell-wall length contractions upon plasmolysis (**d**), lysis (**e**) and in total (**f**) under chemical and genetic perturbations to the outer membrane (ratio with respect to the wild type (WT); $n = 79$, 309, 65, 70, 65, 55, 59 and 50 cells for the wild-type, EDTA, FM 4-64, *imp4213*, Δ*ompA*, +O8, Δ*lpp* and Δ*pal* groups, respectively). *P* values are from a Student's two-sided *t*-test of the difference from the untreated wild-type control. NS, not significant. † (in **e**), the ratio for cells expressing the O8 antigen was very large and not well-defined because the contraction upon lysis for the parental wild-type strain (AB1133) was approximately equal to zero, and adding the antigen markedly increased contraction (Extended Data Fig. 3h). **g**, Outer-membrane stiffness under chemical or genetic perturbations (ratio with respect to the wild type). Data are mean ± s.d. Uncertainty propagated from ε_l and ε_p measurements. **h**, Sorbitol concentration in growth medium during 100-mM oscillatory osmotic shocks with a 2-min period. **i**, Representative cell-wall lengths during shocks in **h** (blue traces; $n = 243$ cells). *A*, amplitude of cell-wall length oscillations. Green shading indicates the period in which EDTA was included. The red curve represents the effective population-averaged length. **j**, Mean amplitude of cell-wall length oscillations during 100-mM oscillatory shock with a 2-min period ($n = 10$ cycles for each measurement). Data are mean ± s.d. *P* values are from a Student's two-sided *t*-test, of the difference from the untreated wild-type control.

that weakened the outer membrane during osmotic shock-based assays also decreased cell stiffness (Fig. 3c, Extended Data Fig. 6c). However, the O8 antigen did not increase cell stiffness (Fig. 3c), despite the strong phenotype we observed in osmotic shock-based assays (Fig. 2d–g, j). We speculate that because the O8 antigen greatly increases the length of LPS, it results in a thick layer that 'pads' the cell against perpendicular indentation forces but can nonetheless bear in-plane loads. Deletion of *lpp* markedly reduced cell stiffness (Fig. 3c). Deletion of *pal* rendered the cells extremely sensitive to indentation such that they lysed when addressed with the AFM cantilever, precluding stiffness determination.

Next, we used a microfluidics-based assay[20,21] to measure the deformation of wild-type cells in response to bending forces before and after treatment with 10 mM EDTA. A perpendicular fluid force was applied to filamentous cells (Fig. 3d), and the bending rigidity was calculated by fitting the deflection of the cell (Fig. 3e) to a mechanical model[21]. Untreated cells deflected less than did EDTA-treated cells, yielding bending rigidities of $5.8 \times 10^{-20} \pm 0.4\,\mathrm{N\,m^2}$ and $2.4 \times 10^{-20} \pm 0.2\,\mathrm{N\,m^2}$, respectively (Fig. 3f), again demonstrating the contribution of LPS to outer-membrane stiffness against various types of load.

Given the profound contribution of the outer membrane to the mechanical properties of the cell envelope, we next explored their contribution to cell physiology. Wild-type *E. coli* cells gradually lysed during moderately large (400 mM) oscillatory osmotic shocks (Fig. 4a, Supplementary Video 4). The chemical perturbations *N*-lauroyl sarcosine, EDTA and FM 4-64 markedly decreased cell survival in the presence of shocks (Fig. 4a, b, Extended Data Fig. 7a, Supplementary Video 4). In the absence of shocks, EDTA and FM 4-64 caused little and no lysis, respectively (Fig. 4a), demonstrating that outer-membrane strength is important for cell survival during mechanical perturbations but not during steady-state growth. Genetic mutations that weakened the outer membrane or disrupted coupling between the outer membrane and the cell wall also markedly decreased cell survival during shocks (Fig. 4b, Extended Data Fig. 7b, d), whereas the presence of the O8 antigen delayed the onset of lysis (Fig. 4b, Extended Data Fig. 7c).

© 2018 Springer Nature Limited. All rights reserved.

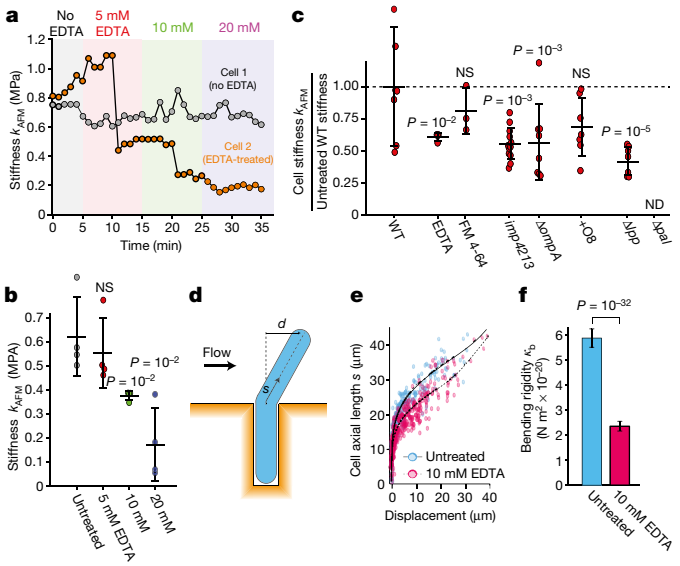


Fig. 3 | The stiffness of turgid cells is dependent on outer-membrane integrity. a, AFM measurements of cell stiffness versus time. One cell was treated with increasing EDTA concentrations (similar data for four treated and two untreated cells). **b**, Mean cell stiffness versus EDTA concentration ($n = 4$ cells for each measurement). Data are mean ± s.d. *P* values are from a Student's two-sided *t*-test of the difference from the untreated wild-type control. **c**, AFM measurements of cell stiffness under chemical or genetic perturbations (relative to the untreated wild type, $n = 7$, 4, 3, 13, 8, 7 and 7 cells for the wild-type, EDTA, FM 4-64, *imp4213*, Δ*ompA*, +O8, Δ*lpp* and Δ*pal* groups, respectively). ND, Δ*pal* cells lysed under AFM. Data are mean ± s.d. *P* values from a Student's two-sided *t*-test of the difference from the untreated wild-type control. **d**, Microfluidic cell-bending assay. **e**, Displacement versus axial length for untreated (blue dots, $n = 300$) and EDTA-treated (pink dots, $n = 367$) cells. Solid and dashed black lines, best fits of the mechanical model used to calculate the bending rigidities. **f**, Mean bending rigidities of untreated and EDTA-treated cells. Error bars, 95% confidence intervals. *P* value calculated from the confidence intervals.

Finally, we hypothesized that the strength of the outer membrane would affect the survival of L-forms, viable wall-less cells that are produced by inhibiting cell-wall biosynthesis[22]. All chemical and genetic perturbations that reduced outer-membrane stiffness also markedly reduced the number of viable L-forms generated after inhibiting wall synthesis (Fig. 4c, Extended Data Fig. 7e). During L-form generation in *imp4213* cells, which yielded no viable L-forms (Fig. 4c), cells always lysed as soon as the cytoplasm blebbed out of the cell wall, which is precisely when turgor pressure is shifted entirely to the outer membrane (Supplementary Video 5), consistent with previous measurements showing that outer membrane integrity affect the dynamics of blebbing and lysis during antibiotic treatment[23]. In summary, the mechanical properties of the outer membrane are critical for cell survival during osmotic fluctuation and L-form proliferation.

Our conclusions revise our understanding of the physiology of both the cell wall and the outer membrane: whereas the former determines cell shape and the latter provides a selective chemical barrier, both are important mechanical elements in Gram-negative bacteria. Our finding that the outer membrane is under negligible or slightly compressive load during steady-state growth (Fig. 2b, Extended Data Fig. 3g) may provide a mechanism for the formation of outer-membrane vesicles, which could bud spontaneously under this circumstance. However, steady-state growth is unlikely to be common for *E. coli*: in the gut, bacteria frequently encounter osmotic fluctuations and dynamic mechanical forces, which will constantly engage the load-bearing nature of the outer membrane. Cell-wall biosynthesis has long been an antibiotic target because of the role of the cell wall in protecting the cell from osmotic lysis, and L-forms have been implicated in antibiotic persistence[24]. Hence, our finding that the outer membrane shares mechanical load with the cell wall, and is therefore critical for L-form proliferation, has

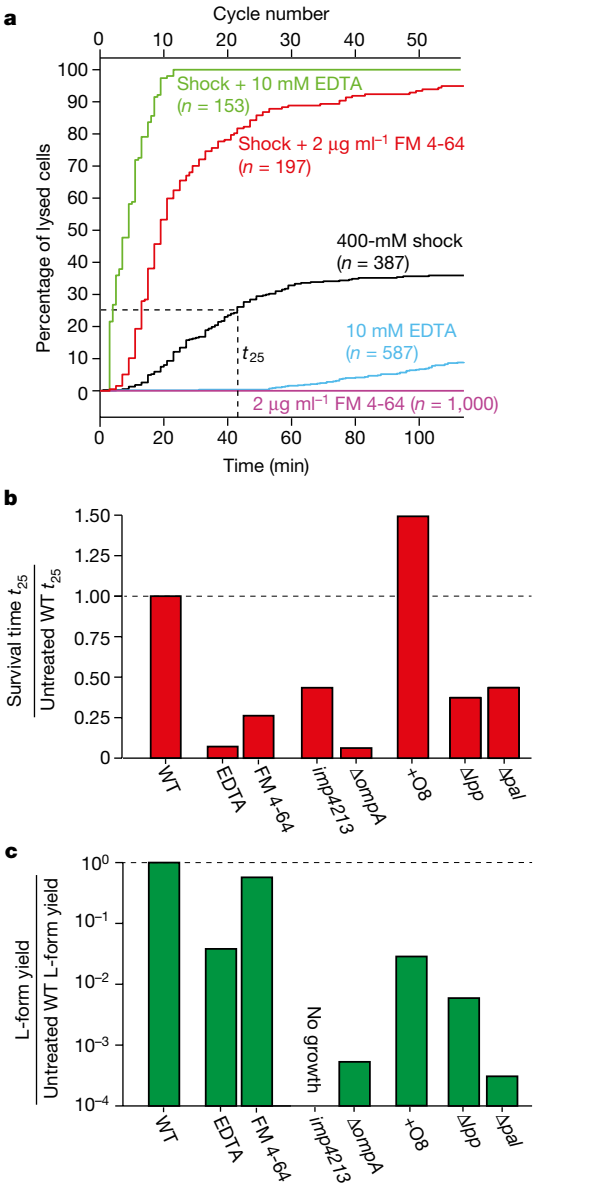


Fig. 4 | Undermining outer-membrane integrity reduces survival during mechanical perturbation and L-form proliferation. a, Cell lysis versus time under chemical perturbation to outer membrane. Shock, 400-mM oscillations with a 2-min period. No cells died in the absence of both oscillatory shock and outer membrane perturbation ($n = 1{,}000$ cells). t_{25}, time at which 25% of cells had lysed. **b**, Time at which 25% of cells had lysed under chemical or genetic perturbations (ratio to the untreated wild type). **c**, Concentration of viable L-forms in overnight cultures under chemical or genetic perturbations to the outer membrane (ratio to the untreated wild type, one experiment).

clear consequences for antibacterial therapy. Further research should elucidate the interdependence between the chemical and mechanical roles of the outer membrane. In particular, it is notable that LPS, which is a potent antigen that stimulates the mammalian immune system, thereby endangering pathogenic bacteria, also provides mechanical integrity to bacteria and therefore protects them.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0344-3.

Received: 10 September 2016; Accepted: 5 June 2018;
Published online 18 July 2018.

© 2018 Springer Nature Limited. All rights reserved.

1. Zgurskaya, H. I., Löpez, C. A. & Gnanakaran, S. Permeability barrier of Gram-negative cell envelopes and approaches to bypass it. *ACS Infect. Dis.* **1**, 512–522 (2015).
2. Martin, H. & Frank, H. Quantitative bausteinanalyse der stützmembran in der zellwand von *Escherichia coli* B. *Z. Naturforsch. B* **17**, 190–196 (1962).
3. Deng, Y., Sun, M. & Shaevitz, J. W. Direct measurement of cell wall stress stiffening and turgor pressure in live bacterial cells. *Phys. Rev. Lett.* **107**, 158101 (2011).
4. Höltje, J. V. Growth of the stress-bearing and shape-maintaining murein sacculus of *Escherichia coli*. *Microbiol. Mol. Biol. Rev.* **62**, 181–203 (1998).
5. Koch, A. L. Biophysics of bacterial walls viewed as stress-bearing fabric. *Microbiol. Rev.* **52**, 337–353 (1988).
6. Herrmann, M., Schneck, E., Gutsmann, T., Brandenburg, K. & Tanaka, M. Bacterial lipopolysaccharides form physically cross-linked, two-dimensional gels in the presence of divalent cations. *Soft Matter* **11**, 6037–6044 (2015).
7. Rassam, P. et al. Supramolecular assemblies underpin turnover of outer membrane proteins in bacteria. *Nature* **523**, 333–336 (2015).
8. Ursell, T. S., Trepagnier, E. H., Huang, K. C. & Theriot, J. A. Analysis of surface protein expression reveals the growth pattern of the Gram-negative outer membrane. *PLOS Comput. Biol.* **8**, e1002680 (2012).
9. Cayley, D. S., Guttman, H. J. & Record, M. T., Jr. Biophysical characterization of changes in amounts and activity of *Escherichia coli* cell and compartment water and turgor pressure in response to osmotic stress. *Biophys. J.* **78**, 1748–1764 (2000).
10. Rojas, E., Theriot, J. A. & Huang, K. C. Response of *Escherichia coli* growth rate to osmotic shock. *Proc. Natl Acad. Sci. USA* **111**, 7807–7812 (2014).
11. de Vries, H. *Die Plasmolytische Studien über die Wand Vacuolen* (G. Bernstein, Berlin, 1885).
12. Howatson, A. M. *Engineering Tables and Data* (Springer Science & Business Media, Berlin, 2012).
13. Tuson, H. H. et al. Measuring the stiffness of bacterial cells from growth rates in hydrogels of tunable elasticity. *Mol. Microbiol.* **84**, 874–891 (2012).
14. Coughlin, R. T., Peterson, A. A., Haug, A., Pownall, H. J. & McGroarty, E. J. A pH titration study on the ionic bridging within lipopolysaccharide aggregates. *Biochim. Biophys. Acta* **821**, 404–412 (1985).
15. Broady, K. W., Rietschel, E. T. & Lüderitz, O. The chemical structure of the lipid A component of lipopolysaccharides from *Vibrio cholerae*. *Eur. J. Biochem.* **115**, 463–469 (1981).
16. Leive, L., Shovlin, V. K. & Mergenhagen, S. E. Physical, chemical, and immunological properties of lipopolysaccharide released from *Escherichia coli* by ethylenediaminetetraacetate. *J. Biol. Chem.* **243**, 6384–6391 (1968).
17. Amro, N. A. et al. High-resolution atomic force microscopy studies of the *Escherichia coli* outer membrane: structural basis for permeability. *Langmuir* **16**, 2789–2796 (2000).
18. Ruiz, N., Falcone, B., Kahne, D. & Silhavy, T. J. Chemical conditionality: a genetic strategy to probe organelle assembly. *Cell* **121**, 307–317 (2005).
19. Sampson, B. A., Misra, R. & Benson, S. A. Identification and characterization of a new gene of *Escherichia coli* K-12 involved in outer membrane permeability. *Genetics* **122**, 491–501 (1989).
20. Amir, A., Babaeipour, F., McIntosh, D. B., Nelson, D. R. & Jun, S. Bending forces plastically deform growing bacterial cell walls. *Proc. Natl Acad. Sci. USA* **111**, 5778–5783 (2014).
21. Auer, G. K. et al. Mechanical genomics identifies diverse modulators of bacterial cell stiffness. *Cell Syst.* **2**, 402–411 (2016).
22. Billings, G. et al. *De novo* morphogenesis in L-forms via geometric control of cell growth. *Mol. Microbiol.* **93**, 883–896 (2014).
23. Yao, Z., Kahne, D. & Kishony, R. Distinct single-cell morphological dynamics under beta-lactam antibiotics. *Mol. Cell* **48**, 705–712 (2012).
24. Kawai, Y., Mickiewicz, K. & Errington, J. Lysozyme counteracts β-lactam antibiotics by promoting the emergence of L-form bacteria. *Cell* **172**, 1038–1049.e10 (2018).

Acknowledgements The authors thank members of the Huang, Theriot, and Weibel laboratories for discussions, K. Amberg-Johnson for assistance with immunoblotting, and T. Silhavy, T. Bernhardt, A. Maurelli, B. Hammer and H. Arjes for feedback, antibodies and strains. This work was supported by National Institutes of Health (NIH) Director's New Innovator Awards DP2OD006466 (to K.C.H.) and DP2OD008735 (to D.B.W.), National Science Foundation (NSF) CAREER Award MCB-1149328 (to K.C.H.), the Stanford Systems Biology Center funded by NIH grant P50 GM107615 (to K.C.H. and J.A.T.), NIH Grant R37-AI036929 (to J.A.T.), the Howard Hughes Medical Institute (to J.A.T.), and NSF Grant DMR-1121288 (to D.B.W.). K.C.H. is a Chan Zuckerberg Biohub Investigator. E.R.R. was supported by a postdoctoral fellowship from the Simbios Center for Physics Based Computation at Stanford University under NIH Grant U54 GM072970. P.D.O. was supported by a postdoctoral fellowship from the Swiss National Science Foundation under Grant P2ELP3_172318. This work was also supported in part by the NSF under Grant PHYS-1066293, the hospitality of the Aspen Center for Physics and by the Allen Discovery Center program through The Paul G. Allen Frontiers Group.

Reviewer information *Nature* thanks J.-F. Collet, W. Margolin, N. Minc and A. Rutenberg for their contribution to the peer review of this work.

Author contributions E.R.R. conceptualized the study. E.R.R., G.B., P.D.O., G.K.A., A.M., F.C., D.B.W., J.A.T. and K.C.H. designed the experiments. E.R.R. and G.B. performed detergent treatments, oscillatory osmotic shock assays and protein release assays. G.B. performed cloning. E.R.R. and P.D.O. performed AFM experiments. G.K.A. performed microfluidic bending assays. E.R.R., G.B. and L.Z. performed L-form experiments. A.M. performed ultra performance liquid chromatography experiments. E.R.R., G.B., P.D.O., A.M. and G.K.A. analysed the data. E.R.R., G.B., G.K.A., P.D.O., A.M., J.A.T. and K.C.H. wrote the manuscript. All authors reviewed the manuscript before submission.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0344-3.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0344-3.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to K.C.H.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Bacterial strains and culture. Bacterial strains used in this study are listed in Supplementary Table 1. Three wild-type *E. coli* genetic backgrounds were used: MG1655, MC4100 and AB1133. We tested the effect of the *imp4213* allele in the MG1655 and MC4100 backgrounds. The Δ*ompA*, Δ*pal*, and Δ*lpp* deletions were tested in the MC4100 background. The O8 antigen was expressed in the AB1133 background (ATM378)[25]. Strain KC427 (MG1655 *imp4213 carB::Tn10*) was constructed using P1 transduction to transfer the *imp4213* mutation from MC4100 to MG1655 cells using NR693 lysate[18]. Bacteria (including non-*E. coli* species) were grown in lysogeny broth (LB), Lennox formulation (5 g l^{-1} NaCl) or minimal medium M9 with 20 mM glucose, overnight in a rotary shaker at 37 °C.

Statistical significance. Under the hypothesis that the outer membrane can bear large mechanical loads, we expected that chemical or genetic perturbations to the outer membrane would result in large changes in the deformation of the cell envelope under a given load. We therefore conservatively calculated the number of cells required to distinguish such a change using a two-sided Student's *t*-test to be $n \approx 44$ to obtain a power of $\beta = 0.9$ with statistical significance $P = 0.05$ and effect size with Cohen's $d = 0.5$.

Imaging in microfluidic devices. Cells were imaged on a Nikon Eclipse Ti-E inverted fluorescence microscope with a 100× (NA 1.40) oil-immersion objective. Images were collected on a DU885 electron-multiplying charge-coupled device camera (Andor) using μManager v.1.4[26]. Cells were maintained at 37 °C during imaging with an active-control environmental chamber (HaisonTech).

Overnight cultures were diluted 100-fold into 1 ml of fresh LB and incubated for 2 h with shaking at 37 °C, except in experiments in minimal medium, for which overnight cultures were diluted into M9 with 20 mM glucose and incubated for 8 h. Cells were imaged in B04A microfluidic perfusion plates (CellASIC) and medium was exchanged using the ONIX microfluidic platform (CellASIC). Plates were loaded with medium pre-warmed to 37 °C. Cells were loaded into the plate, which was incubated at 37 °C, without shaking, for 30 min before imaging. The cell envelope was stained with wheat germ agglutinin–AlexaFluor488 (WGA-AF488, Life Technologies), the fluorescent D-amino acid HADA (gift from the Brun and Vannieuwenzhe Laboratories, University of Indiana), and/or FM 4-64 (Thermo Fisher). For staining of the cell wall, WGA-AF488 or HADA was added to the loading well to a final concentration of 10 μg ml^{-1} or 1 mM, respectively, before loading cells into the imaging chamber. For staining of the outer membrane, FM 4-64 was added to the loading and perfusion medium to a final concentration of 2 μg ml^{-1} and medium containing FM 4-64 was perfused for 3 h before imaging to saturate the polydimethylsiloxane of the microfluidic chamber with the dye. The osmolarity of the growth medium or phosphate-buffered saline (PBS) was modulated with sorbitol (Sigma).

For plasmolysis and lysis and oscillatory osmotic-shock experiments to measure the effect of chemical perturbations (EDTA and FM 4-64 labelling) on cell stiffness, MG1655-based strains were stained with WGA-AF488; non-MG1655 strains stained poorly with WGA-AF488 (data not shown). For plasmolysis and lysis experiments, *E. coli* MC4100- and AB1133-based strains, *Pseudomonas aeruginosa*, and *V. cholerae* were stained with HADA. For oscillatory osmotic shocks of non-MG1655 strains, HADA photobleached too quickly for tracking over many frames (data not shown); hence, we stained and tracked cell envelopes with FM 4-64. As FM 4-64 itself increased the amplitude of oscillations in MG1655 cells (Fig. 2j), ratios between the amplitudes of each FM 4-64-stained mutant and the FM 4-64-stained wild-type parental strain were reported (Fig. 2j).

During plasmolysis and lysis experiments, cells were allowed to grow for 5 min in medium in the imaging chamber before being plasmolysed using medium with 3 M sorbitol and exposed to medium with 3 M sorbitol and 5% *N*-lauroyl sarcosine 5 min later. For plasmolysis and lysis in PBS, cells were loaded into the imaging chamber in LB, which was exchanged for PBS with 70 mM sorbitol (to balance the osmolarity of PBS with LB) for 10 min, then exchanged with PBS and 3 M sorbitol for 5 min, and finally exchanged using PBS with 3 M sorbitol and 5% *N*-lauroyl sarcosine. For plasmolysis and lysis in the presence of EDTA, 10 mM EDTA was included in the plasmolysis and lysis media.

To measure the rest length of the outer membrane, we plasmolysed cells stained with WGA-AF488 and FM 4-64, as described above. We then exchanged the plasmolysis medium with the same medium containing 1 mg ml^{-1} lysozyme (Thermo Fisher). After approximately 15 min, cells began to lose their rod-like shape and lyse. Upon lysis, when released from the constraint of the cell wall, the outer membrane of many cells collapsed into many small vesicles. However, the outer membranes of many other cells maintained their topology and their size could be measured (see below).

When EDTA was applied to cells after plasmolysis (Fig. 2c), for reasons we do not understand, a larger fraction of cells quickly recovered from plasmolysis and swelled. Pre-treatment with FM 4-64 resulted in a similar de-plasmolysis (data not shown), indicating that chemical perturbation of outer membrane integrity is somehow responsible. Cells that de-plasmolysed were not included in population-averaged contraction measurements (Fig. 2d–f).

For oscillatory osmotic shocks, cells were allowed to grow for 5 min in medium in the imaging chamber before being subjected to 100-mM oscillatory osmotic shocks by switching between LB and LB with 100 mM sorbitol. When adding EDTA during oscillatory shocks, two types of experiments were performed. For illustrative purposes, in one experiment cells were left untreated for ten cycles, and then treated with EDTA for ten cycles (Fig. 2h, i). In a complementary set of experiments, EDTA was added during the first ten cycles (but not before); this experiment was used to calculate the amplitude ratio for EDTA treatment (Fig. 2j).

To measure lysis curves (Fig. 4a, Extended Data Fig. 7a–c), cells were loaded into the microfluidic chip as described above, and imaged in LB for 5 min before beginning chemical perturbation and/or oscillatory osmotic shock. A cell was considered to have lysed when it lost phase contrast and, in the case of oscillatory shocks, cell size ceased to oscillate. In the absence of shocks, most strains tested showed no lysis after 2 h of growth in the microfluidic chamber; Δ*pal* and MG1655 *imp4213* cells showed a small degree of lysis (2.0% and 3.5%, respectively) after 2 h of steady-state growth in LB.

Cell-growth tracking and analysis. Custom MATLAB software was written to automatically track cells stained with WGA-AF488 or FM 4-64, as in previous studies[10]. In plasmolysis and lysis experiments in which cells were stained with HADA, the signal-to-noise ratio was not high enough to track cell-wall contours automatically, so cell lengths (immediately before plasmolysis, immediately after plasmolysis and 20 min after detergent treatment) were measured manually in FIJI.

To calculate outer-membrane rest lengths, we used a custom MATLAB algorithm to manually trace cell-envelope contours before plasmolysis, after plasmolysis and after lysozyme-induced lysis. In the first two states, cell length and width were automatically calculated. After lysozyme treatment, we calculated the length of the outer membrane as if it were rod-shaped with a width equal to the width of the cell before plasmolysis, as follows. We first calculated the surface area of the turgid cell envelope, A_{turgid}, from its contour by assuming rotational symmetry around its long axis. After lysis, the outer membrane adopted an amorphous morphology and remained trapped in the microfluidic chamber with height equal to the width of the cell before lysis, w. Therefore, for each cell we calculated the 'rest surface area' of the outer membrane, A_{om}, by first measuring the arc length of the cell contour, s, and the area enclosed within this contour, a, and then applied the equation $A_{om} = 2a + ws$. Finally, to calculate the rest length of the outer membrane, l_{om}, defined as the length of the outer membrane if it had surface area equal to A_{om} and was rod-shaped with a cell width equal to w, we used the equation $l_{om} = l_{turgid} + (A_{turgid} - A_{om})/(\pi w)$. This equation indicates that the difference between the rest length of the outer membrane and the length of the turgid cell is equal to the length of a cylindrical section with area $(A_{turgid} - A_{om})$ and radius $w/2$.

To calculate the amplitude of length oscillations during oscillatory osmotic shocks, cells were tracked using custom MATLAB algorithms. First, cell-wall lengths (l) were automatically tracked (Extended Data Fig. 5f, g), and the elongation rate ($\dot{e} = d(\ln l)/dt$) was calculated for each cell (Extended Data Fig. 5h). The effective population-averaged length was calculated by integrating the population-averaged elongation rate over time[10] (Extended Data Fig. 5i):

$$l_{eff} = \int_{t_1}^{t_2} \dot{e}\,dt + l_0,$$

in which l_0 is the mean initial cell length. The effective population-averaged length was then smoothed with a mean filter with window size equal to the period of oscillation (Extended Data Fig. 5i), and subtracted from the unsmoothed trace to obtain the deviation of the length oscillations around the smoothed trace (Extended Data Fig. 5j). The peak-to-peak amplitude was calculated for each cycle (Extended Data Fig. 5k). The mean amplitude was calculated by averaging the peak-to-peak amplitude over cycles. Uncertainty was estimated as the s.d. of the mean amplitude over cycles.

The time scale of the loss of cytoplasmic contents (cytoplasmic GFP or ribosomal S2-YFP) and the time scale for cell-wall expansion during lysis were calculated using custom MATLAB algorithms. The integrated fluorescence intensity in each cell was computed over time during plasmolysis and lysis experiments. The duration of efflux (or the time scale of expansion during lysis) was measured manually by selecting time points immediately before and after efflux (or lysis).

Outer-membrane release assay. An overnight *E. coli* MG1655 culture was diluted 100-fold into 100 ml fresh LB and incubated with shaking at 37 °C for 2 h. The culture was aliquoted into ten 15-ml tubes and centrifuged (Eppendorf 5804 R) at 3,000*g* for 10 min to pellet cells. The supernatant was decanted and each pellet was resuspended in 1 ml PBS and transferred to ten 1.5-ml tubes. The suspensions were centrifuged (Eppendorf 5415 D) at 15,000*g* to pellet cells. The supernatant

© 2018 Springer Nature Limited. All rights reserved.

was removed and each pellet was resuspended in 100 µl PBS. To these suspensions, 100 µl of detergent or EDTA was added to the specified final concentrations and incubated at room temperature for 10 min. Cells and sacculi were pelleted by centrifuging (Beckman Coulter Optima Max-XP, Rotor TL-100) at 100,000 r.p.m. for 10 min at 4 °C. The supernatant was removed for analysis. LPS was detected in serial dilutions of the supernatant using the Pierce *Limulus* Amebocyte Lysate (LAL) assay (ThermoFisher). After performing this reaction, absorbance at 410 nm was measured using an M200 96-well plate reader (Tecan). We found that detergent above 0.05% interfered with the LAL reaction (data not shown) and therefore we diluted supernatants from the detergent-treated samples 1,000-fold into 0.05% detergent before measuring LPS.

Outer-membrane proteins in the supernatant were detected with immunoblotting using an anti-OMP antiserum generously provided by the Thomas Silhavy Laboratory at Princeton University. Supernatant was added to 1× NuPAGE LDS sample buffer (Invitrogen). Proteins were separated by electrophoresis on a 4–12% Bis-Tris gel (Invitrogen) and transferred to a nitrocellulose membrane. After blocking, membranes were probed with 1:30,000 anti-OMP antiserum and 1:10,000 donkey anti-rabbit 800CW (LiCor Biosciences). Fluorescence antibody-bound proteins were detected with an Odyssey Imager (LiCor Biosciences).

Analysis of peptidoglycan composition. Ultra performance liquid chromatography (UPLC) samples were prepared as previously described[27]. Overnight cultures of *E. coli* MG1655 were diluted 1:200 in 250 ml LB and grown at 37 °C to an $OD_{600\,nm}$ of 0.7. Cultures were centrifuged at 5,000*g* for 10 min at room temperature and the resulting pellet was suspended in 3 ml of either LB, LB with 3 M sorbitol, LB with 3 M sorbitol and 10 mM EDTA or LB with 3 M sorbitol and 5% *N*-lauroyl sarcosine. All samples that contained 3 M sorbitol in the resuspension solution were allowed to sit at room temperature for 20 min after resuspension. Cell suspensions were then lysed by boiling in SDS for 3 h. Lysed cell suspensions were ultracentrifuged at 400,000*g* to purify sacculi, which were digested with muramidase into muropeptides. Samples were pH-adjusted and injected onto a Waters H Class UPLC system equipped with a BEH C18 1.7-µm column and eluted with sodium phosphate buffers. Peaks were quantified and identified as particular muropeptide species from their elution times[28], from which the crosslinking density and glycan-strand length were calculated as previously described[29].

Derivation of a mechanical model of the cell envelope. Cell-wall length contractions upon plasmolysis (ε_p), upon lysis (ε_l) and total contraction (ε_t) were defined as: $\varepsilon_p = (l_1 - l_2)/l_2 \times 100$, $\varepsilon_l = (l_2 - l_{cw})/l_{cw} \times 100$ and $\varepsilon_t = (l_1 - l_{cw})/l_{cw} \times 100$, respectively, in which l_1 is the length before plasmolysis, l_2 is the length after plasmolysis, and l_{cw} is the length after lysis. We modelled the cell wall–outer membrane complex as two linear springs in parallel with spring constants k_{cw} and k_{om}, respectively, and rest lengths l_{cw} and l_{om}, respectively (Fig. 1a). Although the mechanical properties of the cell envelope have been measured to be nonlinear by atomic force microscopy[3], envelope contraction upon hyperosmotic shock was approximately linear with respect to shock magnitude up to the maximum contraction caused by large shocks (Extended Data Fig. 1), motivating our use of a linear spring-based model for estimating outer-membrane stiffness. Initially, the springs are subjected to an applied force, F, representing turgor pressure, and have initial length l_1. When the force is released (when the cell is plasmolysed), the springs relax to length l_2. The force balance equation is:

$$k_{cw}(l_2 - l_{cw}) + k_{om}(l_2 - l_{om}) = 0 \quad (2)$$

We substitute for l_2 using the definition of ε_l, which yields:

$$k_{om} = \frac{\varepsilon_l}{\frac{l_{om}}{l_{cw}} - (\varepsilon_l + 1)} k_{cw} \quad (3)$$

Using the experimental finding that $l_{om} = l_l = l_{cw}(1 + \varepsilon_t)$ (Fig. 2b, Extended Data Fig. 3g) and substituting the identity $\varepsilon_t = (1 + \varepsilon_p)(1 + \varepsilon_l) - 1$ yields $l_{om} = l_{cw}(1 + \varepsilon_p)(1 + \varepsilon_l)$. Finally, substituting this expression into equation (3) yields equation (1) of the main text.

AFM measurements of cell stiffness. The day before AFM, poly-L-lysine-coated glass slides were prepared by boiling 22-mm square cover glasses in 2% Micro-90 detergent. Cover glasses were rinsed in ethanol and baked in a dry oven at 80 °C for ≥1 h. Cover glasses were glued onto microscope slides using super glue.

On the day of measurement, 10 µl of 0.01% poly-L-lysine were spotted onto the centre of the cover glass and dried for 1 h. A silicone well (Grace Biolabs) was affixed to the cover glass around the poly-L-lysine spot. An overnight culture of *E. coli* was diluted 100-fold into LB and incubated at 37 °C for 1.5 h. One millilitre of culture was spun down in a table-top microcentrifuge for 1 min at 15,000*g* and resuspended in 1 ml de-ionized water. This suspension was then spun down and concentrated in 100 µl de-ionized water. This suspension was added to the well on the poly-L-lysine-coated slide. The slide was incubated at room temperature for 15 min. The well was rinsed with de-ionized water three times to remove cells that were not stuck to the surface. The well was then immersed in 100 µl M9 with 20 mM glucose.

AFM experiments were performed on a Bioscope Resolve (Bruker) with MLCT-BIO-E (Bruker) cantilevers with a nominal spring constant of 0.1 N m^{-1} (for EDTA experiments) or PFQNM-LC (Bruker) cantilevers (0.08 N m^{-1}). Cantilevers were calibrated using the thermal noise method[30]. We measured the deflection sensitivity (nm V^{-1}) from the slope of a force curve recorded on a glass surface. The resonance frequency and q-factor of the cantilever were determined by a Lorentzian fit of the power spectral density of the cantilever recorded using the thermal tune option in the Nanoscope Analysis v.1.8 (Bruker) software. From these parameters, the spring constant was calculated.

A polydimethylsiloxane calibration sample with known stiffness of 3.5 MPa was measured using force volume imaging mode in an area of 1 × 1 µm at a resolution of 32 × 32 pixels at 0.5-Hz ramp rate and 600-nm ramp size (Extended Data Fig. 6d). Force curves were recorded to a maximal deflection error of 5 nm, and analysis of the force curves was accomplished using Nanoscope Analysis v.1.8. For analysis of the polydimethylsiloxane stiffness measured with PFQNM-LC cantilevers, the baseline correction of the force curve was performed using 10–80% of the ramp size, with a tip radius set to 65 nm, the tip half angle to 18°, and Poisson's ratio of 0.3. The approach direction of the force curve was used to analyse the stiffness, assuming a Hertz contact model as a modulus fit model for the interaction of a sphere with a surface, resulting in a mean stiffness of 3.41 ± 0.73 MPa (Extended Data Fig. 6d). These parameters were then used to analyse force curves recorded on immobilized *E. coli* with a PFQNM-LC cantilever. For measurements with MLCT-BIO-E cantilevers, the tip radius was set to 20 nm.

For each cell measured, an image of the immobilized cell was recorded to confirm firm attachment and to position the cantilever at the centre of the cell before recording force curves. Three consecutive force curves with z-range of 800 nm were recorded at a rate of 0.5 Hz to a maximum deflection of 5 nm. Force curves were further analysed using NanoScope Analysis v.1.8. For each cell, the mean stiffness was calculated from the three force curves. In all cases, individual slides or dishes were measured for no more than 1 h.

To compare our reported cell stiffness values (akin to Young's modulus of the cell as if it was a single bulk material, previously shown to be dependent on envelope stiffness[3]) to previous measurements[3], we used the PDMS calibration to compute the relationship between stiffness values in MPa and in N m^{-1} using the Nanoscope analysis software. The previously reported cell stiffness value[3] (0.017 N m^{-1}) converts to a Young's modulus value of 0.5 MPa for our experimental methodology, which compares very favourably with our data in Fig. 3b.

For measuring the effect of FM 4-64 on cell stiffness, single force curves were taken every minute. After several minutes, a concentrated solution (4 µg ml^{-1}) of FM 4-64 was added to a final concentration of 2 µg ml^{-1}.

For measuring the effect of EDTA on cell stiffness, instead of supported coverglass slides, we adhered cells to glass-bottom dishes (Ted Pella, 14026) that had been spotted with poly-L-lysine as described above. After cells adhered, dishes were immersed in 4 ml M9 with glucose medium. For each cell, single force curves were recorded at 1-min intervals. After 5–10 min, the EDTA concentration was increased stepwise every 10 min to final concentrations of 5 mM, 10 mM, and 20 mM by adding 1 ml of appropriately concentrated stock EDTA solution.

Microfluidic assay of bending stiffness. The bending rigidity of *E. coli* cells was measured using a microfluidic assay essentially as described[20,21]. *E. coli* MG1655 was transformed with a plasmid (pDB192) containing *sulA* under an isopropyl β-D-1-thiogalactopyranoside (IPTG)-inducible promoter. Cells were grown overnight in 2 ml LB containing 30 µg ml^{-1} kanamycin and 50 µg ml^{-1} ampicillin. IPTG was added to the medium to induce *sulA* throughout cell growth in the microfluidic flow chamber. Deflection of cells under fluid flow was monitored on a Zeiss Axiovert 100 microscope (Zeiss) equipped with a 60× oil objective. Images were collected with an Andor iXon 3 EMCCD (Andor) using µManager v.1.4[26]. Deflection of the cells was determined using a custom Igor Pro (WaveMetrics) image-analysis algorithm.

Quantification of L-form viability, physiology and growth. Overnight cultures of the appropriate strains were diluted 1:100 into LFLB (LB supplemented with 3.6% sucrose and 10 mM $MgSO_4$). Cultures were incubated at 37 °C for 1 h, at which point cefsulodin was added to a final concentration of 60 µg ml^{-1}. As appropriate, chemical treatments (EDTA and FM 4-64) were also added at this point. Cells were incubated for 12 h with shaking at 30 °C. Five microlitres of serial tenfold dilutions were plated on LFLB with cefsulodin and LB with cefsulodin plates. Plates were incubated at 30 °C for 24 h. Plates were imaged with a Canon EOS Rebel T5i DSLR camera and colony-forming units per ml were counted manually. MC4100-based L-forms formed smaller colonies that were not visible with the DSLR camera, and thus spots from the plates were excised and imaged with an inverted microscope (Nikon Eclipse Ti-E) at 10× magnification. For each L-form culture except AB1133, spotting onto LB with cefsulodin plates reduced yield by at least an order of magnitude relative to LFLB with cefsulodin (data not shown).

© 2018 Springer Nature Limited. All rights reserved.

L-forms produce LPS at levels approximately equal to untreated cells (6.6×10^{-6} and 7.2×10^{-6} endotoxin units per cell for untreated cells and L-forms, respectively). To measure this, 1 ml of overnight culture was washed twice in PBS. *N*-lauroyl sarcosine was added to a final concentration of 5%, and cell suspensions were incubated at 90 °C for 10 min to lyse cells and dissolve their membranes. LPS was detected in tenfold serial dilutions of this solution using the Pierce *Limulus* Amebocyte Lysate assay (Thermo Fisher).

To image L-form generation from *imp4213* cells, a 5-ml culture was grown overnight in LB at 30 °C. The overnight culture was diluted 100-fold into LB and incubated at 30 °C for 2 h. Five microlitres of this culture were spotted onto an LFLB with cefsulodin agarose pad and imaged every minute at 30 °C.

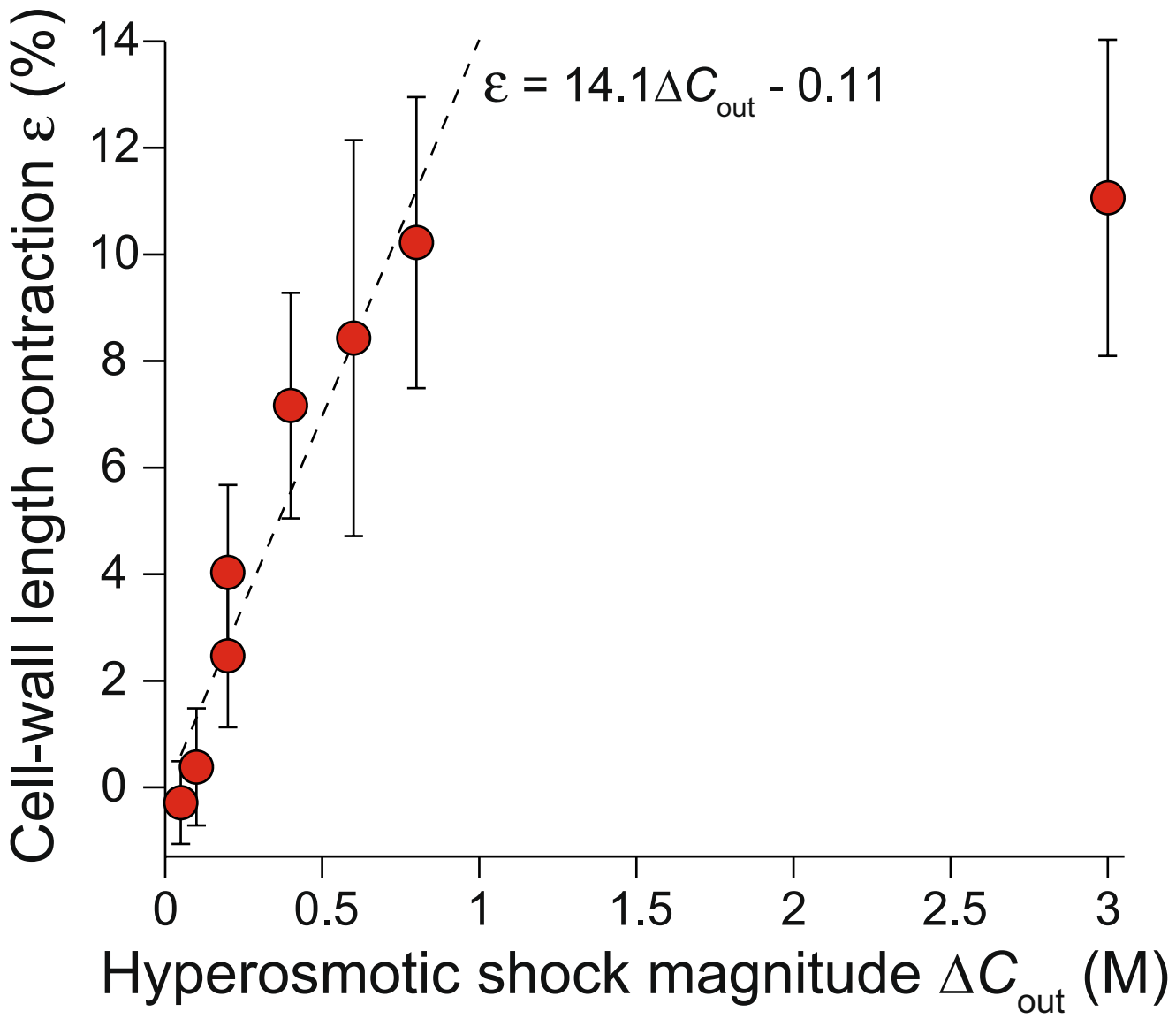
Code availability. All custom code used for the current study is available from the corresponding author on reasonable request.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. The datasets generated during the current study are available from the corresponding author on reasonable request.

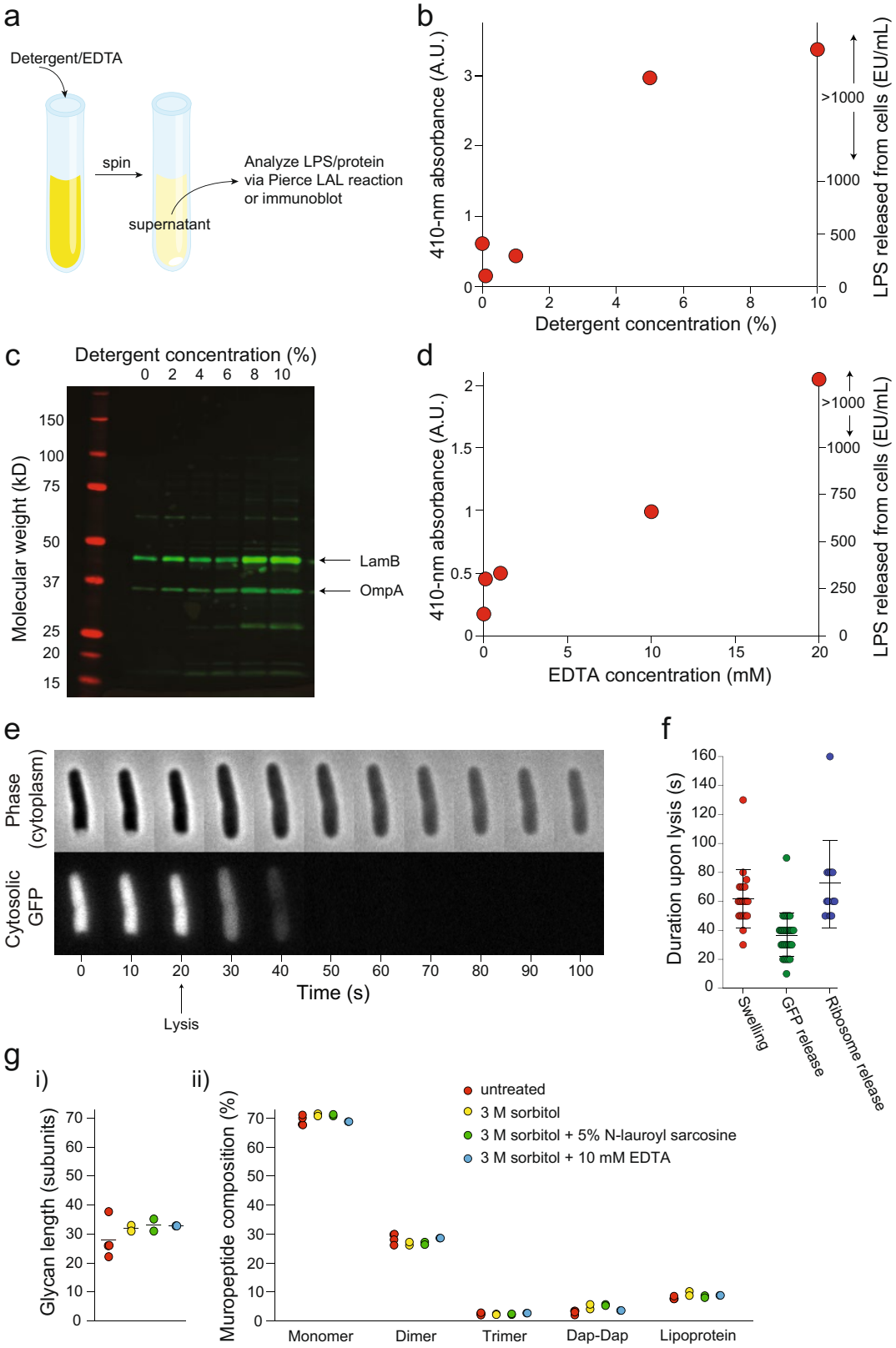
25. Rick, P. D., Hubbard, G. L. & Barr, K. Role of the *rfe* gene in the synthesis of the O8 antigen in *Escherichia coli* K-12. *J. Bacteriol.* **176**, 2877–2884 (1994).
26. Edelstein, A., Amodaj, N., Hoover, K., Vale, R. & Stuurman, N. Computer control of microscopes using μManager. *Curr. Protoc. Mol. Biol.* **92**, 14.20.1–14.20.17 (2010).
27. Desmarais, S. M. et al. High-throughput, highly sensitive analyses of bacterial morphogenesis using ultra performance liquid chromatography. *J. Biol. Chem.* **290**, 31090–31100 (2015).
28. Glauner, B., Höltje, J. V. & Schwarz, U. The composition of the murein of *Escherichia coli*. *J. Biol. Chem.* **263**, 10088–10095 (1988).
29. Glauner, B. Separation and quantification of muropeptides with high-performance liquid chromatography. *Anal. Biochem.* **172**, 451–464 (1988).
30. Hutter, J. L. & Bechhoefer, J. Calibration of atomic-force microscope tips. *Rev. Sci. Instrum.* **64**, 1868–1873 (1993).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 1 | Cell-wall deformation is approximately linear with respect to hyperosmotic shock over a large range. Population-averaged contraction of cell-wall lengths versus hyperosmotic shock magnitude ($n = 92, 73, 53, 71, 58, 11, 31, 47$ cells). Data are mean ± s.d. The dotted line is the linear best fit for experimental data for shocks with magnitude ≤ 800 mM. The plateau after 800 mM demonstrates that the cell envelope has reached its minimum length upon large, 3 M hyperosmotic shocks.

© 2018 Springer Nature Limited. All rights reserved.

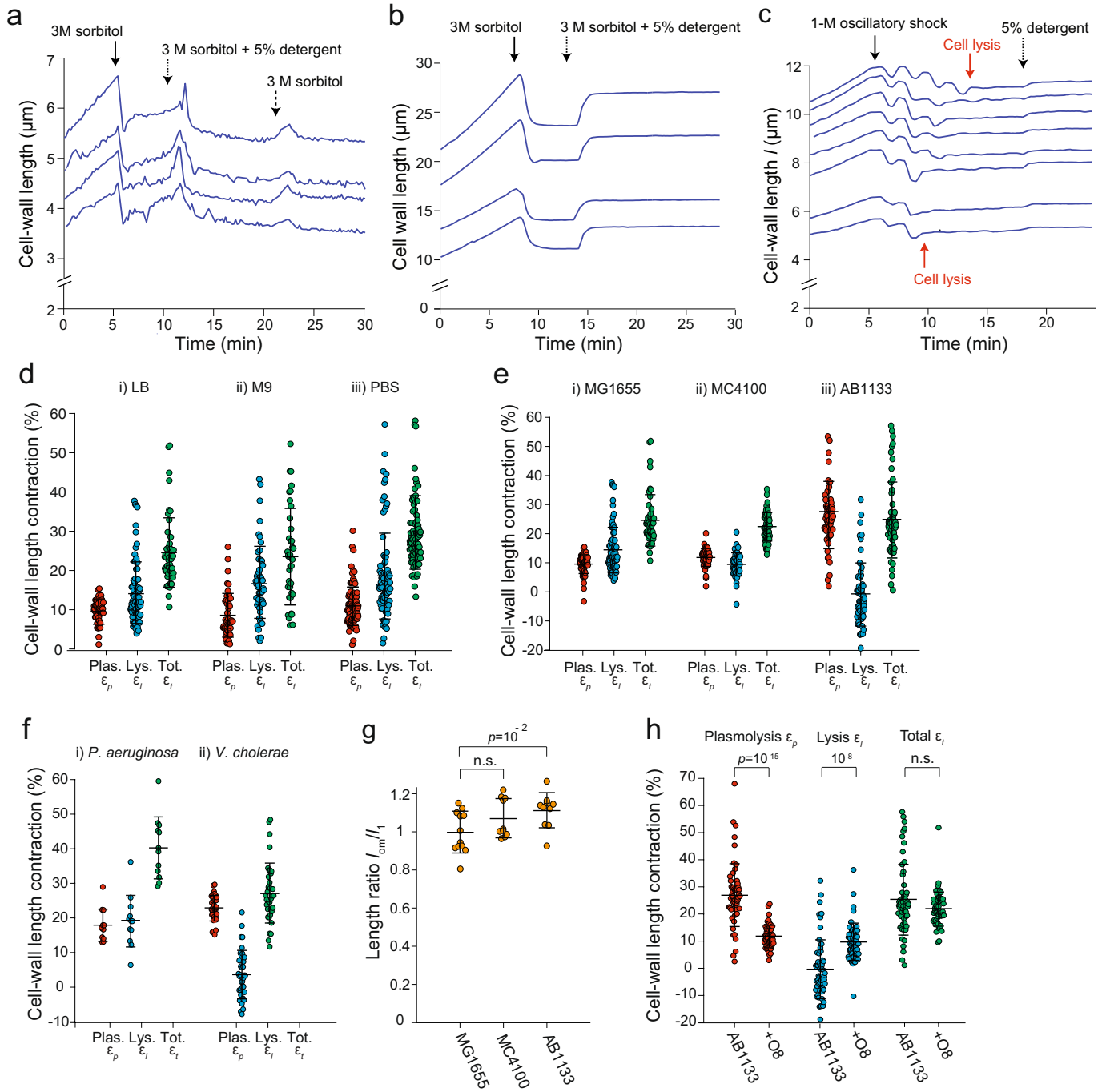


Extended Data Fig. 2 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 2 | Detergent or EDTA treatment causes dissolution of the outer membrane. a, Description of the outer-membrane release assay used to measure the effect of detergent and EDTA on the outer membrane. See Methods section 'Outer membrane release assay' for a more detailed description. **b**, Absorbance of 410-nm light (left axis) in the supernatant from detergent-treated cell suspensions after performing the LAL LPS detection assay. Absorbance is correlated with the amount of LPS in the sample, and is linear below 1,000 EU ml^{-1} (right axis). This experiment was performed once. **c**, Immunoblot with an antibody cross-reactive with several outer-membrane proteins (for example, LamB and OmpA) of the supernatant of cells treated with various concentrations of *N*-lauroyl sarcosine (detergent). This experiment was performed once. **d**, Absorbance of 410-nm light (left axis) in the supernatant from EDTA-treated cell suspensions after performing the LAL LPS detection assay. Absorbance is correlated with the amount of LPS in the sample, and is linear below 1,000 EU ml^{-1} (right axis). This experiment was performed once. **e**, Montage of a representative plasmolysed cell expressing cytosolic GFP lysing upon detergent treatment. Cell expansion corresponds to the time of GFP disappearance. Data are mean ± s.d. from 32 cells in one experiment. **f**, Duration of swelling (left), of release of cytosolic GFP (centre), and of release of ribosomes upon lysis of plasmolysed cells ($n = 22$, 32, 12 cells, respectively; one experiment each). Error bars, ±1 s.d. The duration of swelling is approximately equal to the duration of the release of cytoplasmic contents. **g**, Ultra performance liquid chromatography of peptidoglycan composition of cell walls purified from untreated (red), sorbitol-treated (yellow), detergent-treated (green), and EDTA-treated (blue) cells. Mean glycan strain lengths (**i**) and percentage of various peptidoglycan subunit eluents (**ii**) are the same across all treatments. Dap–Dap, Dimers formed with double diaminopimelic acid bonds as opposed to diaminopimelic acid–D-alanine bonds; lipoprotein, subunits bound to Lpp. $n = 4$, 2, 2 and 1 sacculi preparations for untreated, sorbitol-treated, detergent-treated and EDTA-treated cells, respectively.

© 2018 Springer Nature Limited. All rights reserved.

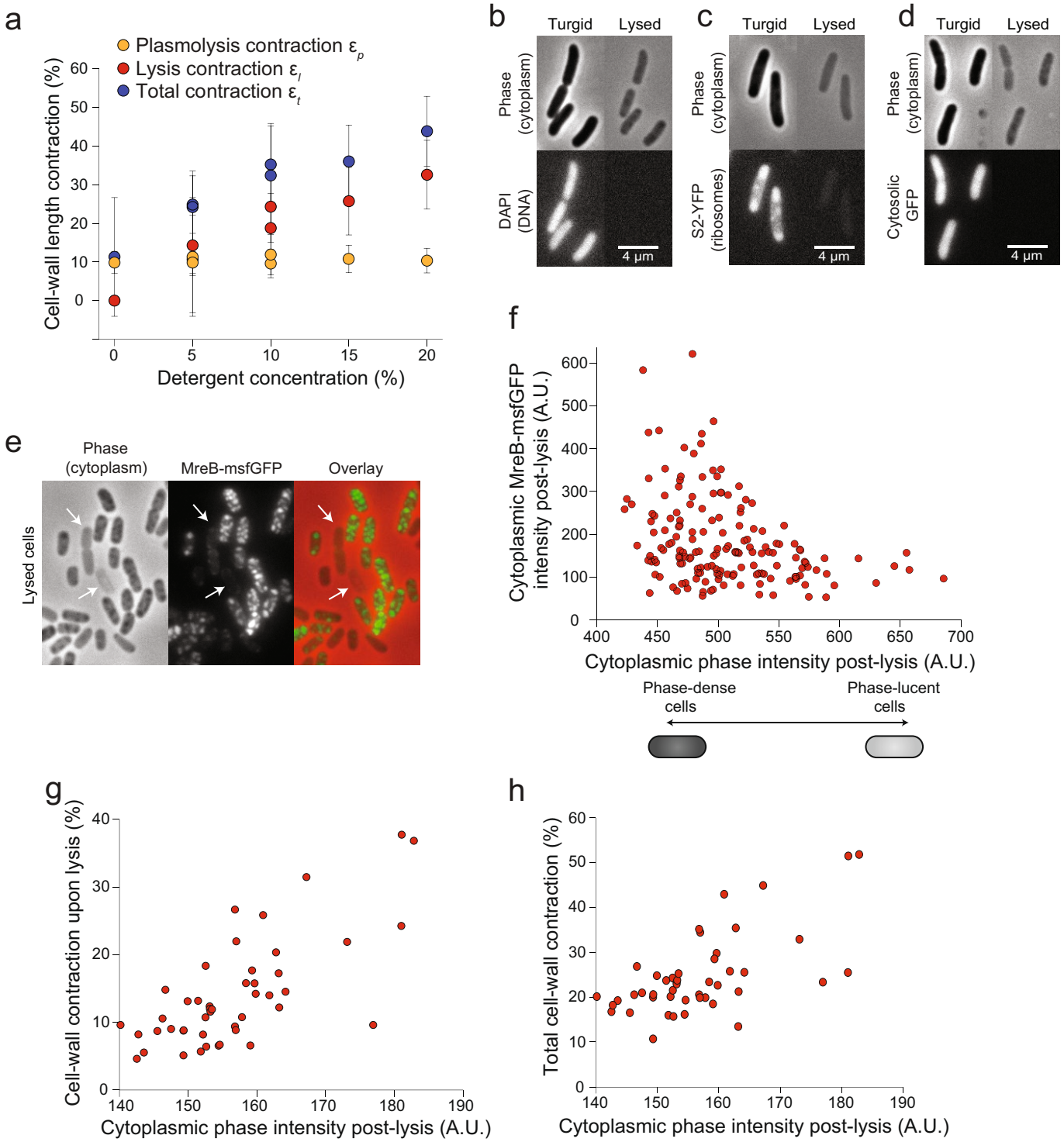


Extended Data Fig. 3 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

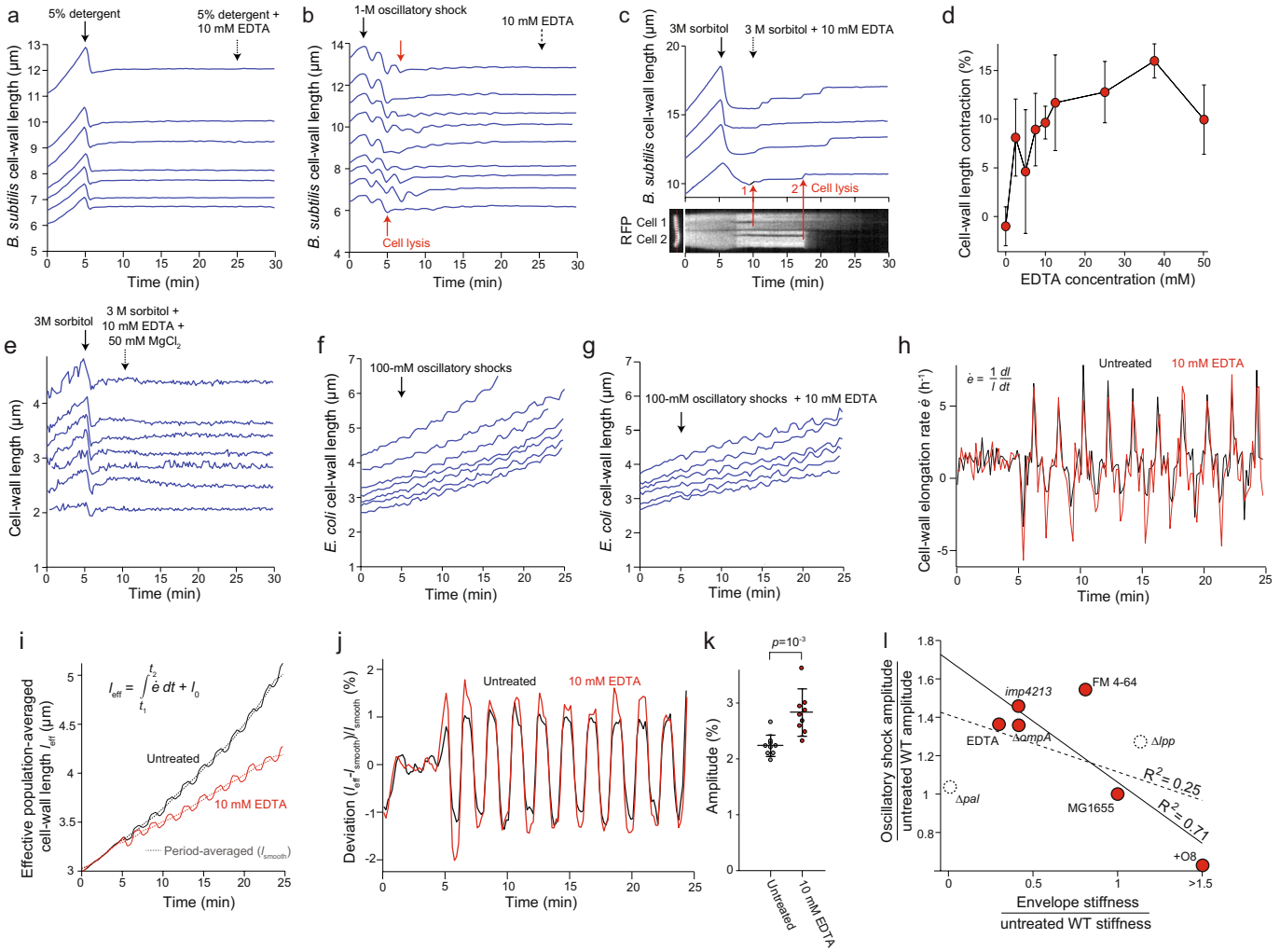
Extended Data Fig. 3 | Detergent treatment causes contraction of the cell wall across a range of conditions and organisms. a, Length of the cell wall versus time during hyperosmotic shock (3 M sorbitol, solid arrow) and subsequent treatment with detergent (5% *N*-lauroyl sarcosine, dotted arrow), and washout of the detergent (switching back to 3 M sorbitol, dashed arrow) for four representative cells ($n = 260$ cells). The cells briefly swelled upon washout, then relaxed to the pre-washout rest length. We speculate that the brief swelling is caused by a hypoosmotic shock during washout (lowering the concentration of detergent). That is, we propose that the detergent has a slow diffusion constant through the cell wall, and therefore washing it out imposes a transient hypoosmotic shock, before detergent within the cell diffuses out of it. When it does, the cell wall relaxes again to its rest state, demonstrating that cell-wall contraction upon detergent treatment is not due to compaction by the detergent. **b**, Lengths of the cell wall versus time during hyperosmotic shock (3 M sorbitol, solid arrow) and subsequent addition of detergent (5% *N*-lauroyl sarcosine, dotted arrow) for four representative *B. subtilis* cell chains ($n = 112$ cell chains). Rather than contract upon detergent treatment, the cell walls re-extend. We hypothesize that the cell wall is in a highly compressed state after 3-M hyperosmotic shock because the cytoplasm is exerting an inward force via connections between the plasma membrane and cell wall. When the plasma membrane is dissolved with detergent, this force is relieved and the cell wall extends to its rest length. **c**, Lengths of the cell wall of *B. subtilis* cell chains versus time during a 1-M oscillatory osmotic shock, which caused cell lysis (for example, red arrows), and treatment with detergent (5% *N*-lauroyl sarcosine, dotted arrow; $n = 93$ cell chains). The cell walls expanded only slightly upon detergent addition, indicating that it is lysis rather than detergent treatment that causes the re-extension in (**b**). **d**, Contraction upon plasmolysis, subsequent lysis and total contraction for three environmental conditions: i) LB ($n = 79$, 56 and 56 cells, respectively), ii) M9 ($n = 46$, 52 and 38 cells, respectively) and iii) cultured in LB and transferred to PBS before plasmolysis ($n = 95$, 86 and 94 cells, respectively). Error bars indicate ± 1 s.d. **e**, Contraction upon plasmolysis, subsequent lysis and total contraction for three wild-type *E. coli* strains: i) MG1655 ($n = 79$, 56 and 56 cells, respectively), ii) MC4100 ($n = 48$, 48 and 48, respectively) and iii) AB1133 ($n = 56$, 56 and 56 cells, respectively). Error bars indicate ± 1 s.d. **f**, Contraction upon plasmolysis, subsequent lysis, and total contraction for i) *P. aeruginosa* ($n = 12$, 12 and 12 cells, respectively) and ii) *V. cholerae* ($n = 36$, 36 and 36 cells, respectively). Error bars indicate ± 1 s.d. **g**, The mean ratio between the rest length of the outer membrane, l_{om}, and the length of the cell envelope in the fully turgid state, l_1, for three wild-type *E. coli* strains ($n = 12$, 10 and 10 cells for MG155, MC4100 and AB1133, respectively). The rest length of the outer membrane is approximately equal to the length of the fully turgid cell for all three strains. Error bars indicate ± 1 s.d. The *P* value was calculated using a Student's two-sided *t*-test. **h**, Mean contraction upon plasmolysis, subsequent lysis, and total contraction for wild-type AB1133 ($n = 56$ cells) and an isogenic strain expressing the O8 antigen ($n = 55$ cells). The ratio for cells expressing the O8 antigen is very large because the contraction upon lysis for the parental wild-type strain (AB1133) was close to zero, and adding the antigen markedly increased contraction. Error bars indicate ± 1 s.d. *P* values were calculated using a Student's two-sided *t*-test.

© 2018 Springer Nature Limited. All rights reserved.



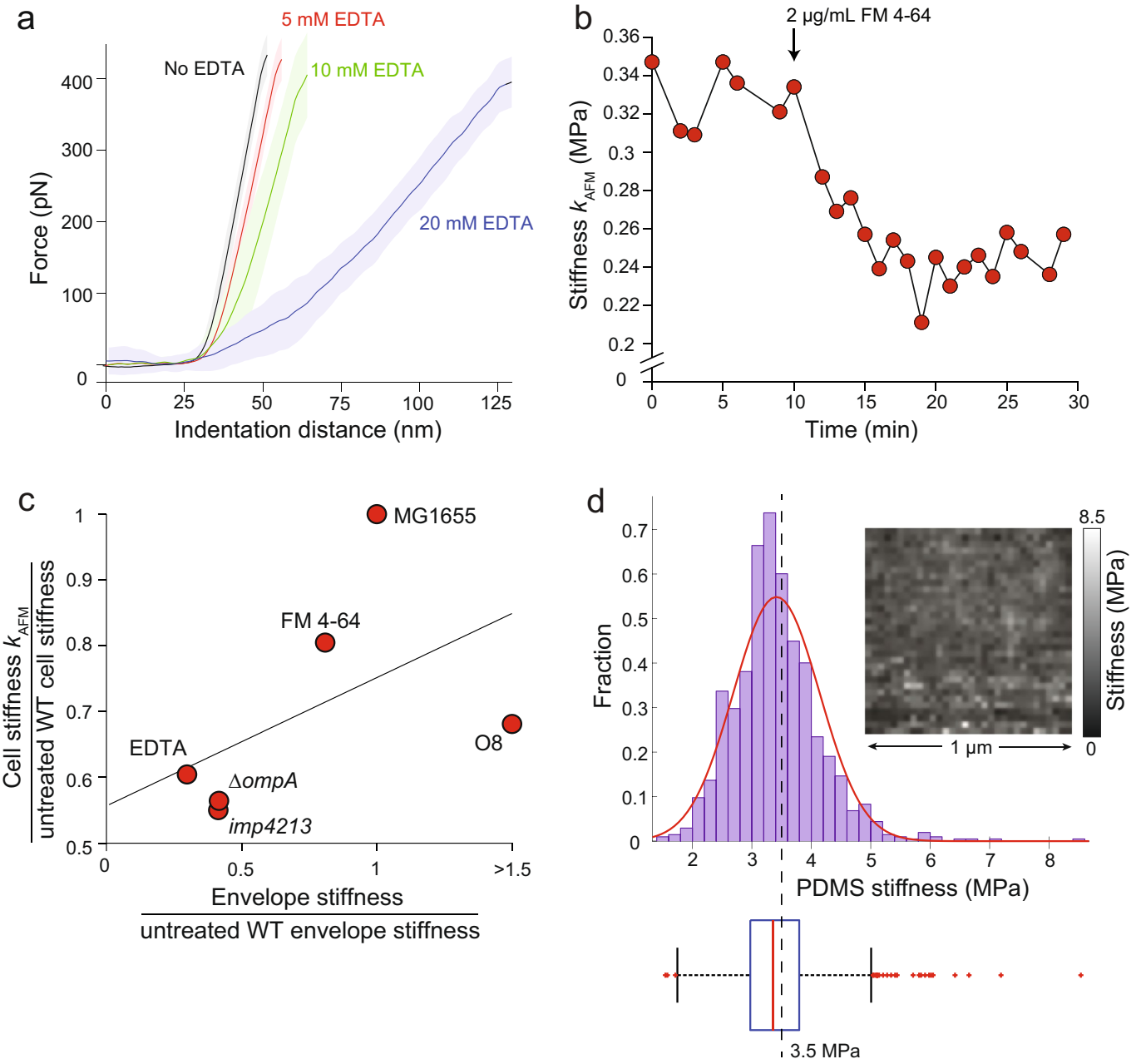
Extended Data Fig. 4 | Impermeable molecules within the cell wall after lysis cause residual turgor pressure. a, Increasing detergent concentration caused an approximately proportional increase in the mean cell-wall contraction upon lysis and the mean total contraction in *E. coli* MG1655 cells. Each point represents one experiment. The number of cells for each experiment is given in Supplementary Table 2. **b**, Representative DAPI-stained *E. coli* cells before (left) and after (right) lysis shown in phase contrast and epifluorescence, demonstrating that DNA is not retained in the lysed cells. The experiment was performed once. **c**, Representative *E. coli* cells expressing a fluorescent protein fusion to the S2 ribosomal protein before (left) and after (right) lysis shown in phase contrast and epifluorescence, demonstrating that ribosomes are not retained in lysed cells. The experiment was performed once. **d**, Representative *E. coli* cells expressing cytosolic GFP before (left) and after (right) lysis shown in phase contrast and epifluorescence, demonstrating that GFP is not retained in lysed cells. The experiment was performed once. **e**, Representative *E. coli* cells expressing a fluorescently tagged version of MreB after lysis shown in phase contrast (left), epifluorescence (centre) and in overlay (right), demonstrating that MreB is retained within most lysed cells, but that cells with weak phase density after lysis retain low levels of MreB (arrows). **f**, Cumulative fluorescence intensity of MreB–sfGFP versus the average phase-contrast intensity within the cell after lysis ($n = 162$ cells). Cells with higher phase density have lower intensity. **g**, Cell-wall contraction upon lysis versus average phase-contrast intensity within the cell after lysis ($n = 46$ cells). **h**, Total contraction during plasmolysis and lysis versus average phase-contrast intensity within the cell after lysis ($n = 46$ cells).

© 2018 Springer Nature Limited. All rights reserved.



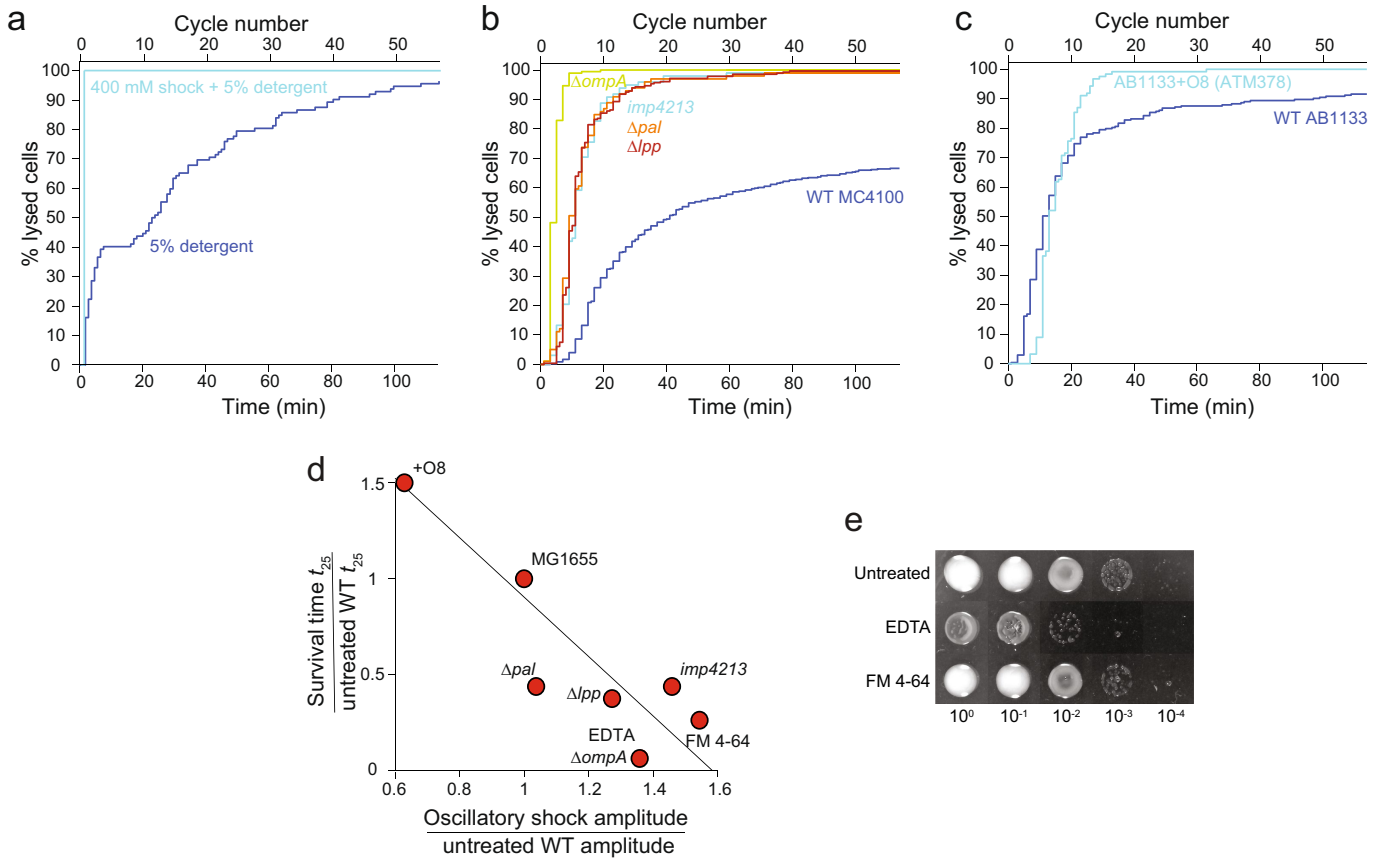
Extended Data Fig. 5 | EDTA weakens the *E. coli* cell envelope. a, Length of the cell wall versus time for 7 representative *B. subtilis* cell chains during treatment with detergent with treatment using detergent and 10 mM EDTA ($n = 68$ cell chains). Although detergent caused lysis, subsequent addition of EDTA did not affect cell-wall rest length. **b**, Lengths of the cell wall of *B. subtilis* cell chains versus time during 1-M oscillatory osmotic shocks, which caused cell lysis (for example, red arrows), with treatment using 10 mM EDTA (dashed arrow; $n = 127$ cell chains). EDTA did not affect the rest length of the cell walls. **c**, Top, length of the *B. subtilis* cell wall versus time through a hyperosmotic shock (3 M sorbitol, solid arrow) and subsequent treatment with 10 mM EDTA (dotted arrow) for four representative cell chains ($n = 61$ cell chains). Cell-wall length did not decrease after detergent, as it did for *E. coli*. Bottom, micrograph of a two-cell chain expressing cytosolic (strain HA405, left) and a kymograph showing fluorescence intensity along the dotted red line during the experiment in the top graph. The cell chain in the kymograph corresponds to the bottom-most length trace in the top graph. Red arrows demonstrate that the discrete increases in length observed after EDTA treatment correspond to cell-lysis events, when fluorescence within the cells begin to decrease. **d**, Population-averaged *E. coli* cell-wall length contraction upon EDTA application after plasmolysis increased with increasing concentration of EDTA ($n = 131, 193, 225, 138, 81, 94, 36, 28, 72$ cells, respectively). Error bars indicate ±1 s.d. **e**, Length of the cell wall versus time during hyperosmotic shock (3 M sorbitol, solid arrow) and subsequent treatment with 10 mM EDTA and 50 mM $MgCl_2$ (dotted arrow) for representative *E. coli* cells ($n = 91$ cells). **f**, Length of the cell walls of representative *E. coli* cells during 100-mM oscillatory shocks with 2-min period ($n = 243$ cells). **g**, Length of the cell walls of representative *E. coli* cells during a 100-mM oscillatory shock with 2-min period and 10 mM EDTA ($n = 284$ cells). **h**, Population-averaged elongation rate of the *E. coli* cell wall during 100-mM oscillatory shocks with 2-min period for untreated (black line) and 10 mM EDTA-treated cells ($n = 284$ cells). **i**, Effective population-averaged cell length (l_{eff}), calculated by integrating the population averaged elongation rate in **h** during 100-mM oscillatory shocks with 2-min period for untreated (black line) and 10 mM EDTA-treated cells ($n = 284$ cells). Dotted lines are the respective time-averaged l_{eff} using a rolling-window averaging filter with a 2-min window (equal to the period of oscillations). **j**, Deviation of the effective population-averaged length in **i** from the respective time-averaged trace. **k**, The mean amplitude of oscillation was found by averaging the peak-to-peak amplitude in **j** over cycles ($n = 10$ cycles). Error bars indicate ±1 s.d. The *P* value was calculated using a Student's two-sided *t*-test. **l**, Amplitude of cell-wall length oscillations (ratio with respect to the untreated wild type; Fig. 2j) versus cell-wall stiffness calculated from plasmolysis–lysis experiments (ratio with respect to the untreated wild type; Fig. 2g). Solid line, linear best fit for only perturbations to the outer membrane (red circles; linear regression: $R^2 = 0.71$, $F = 9.7$, $P = 0.0356$). Dashed line, best fit when additionally considering perturbations to protein linkages between the outer membrane and cell wall (dashed circles; linear regression: $R^2 = 0.25$, $F = 1.4$, not significantly different from horizontal). For the O8-expressing strain, we conservatively used a stiffness ratio of 1.5 for the fits.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 6 | FM 4-64 softens *E. coli* cells. a, Cantilever force versus indentation distance during successively increasing EDTA concentrations, as in Fig. 3a. Lines indicate the average force–distance curves taken during the period in which the cell was treated with the given concentration of EDTA. Shaded areas indicate ±1 s.d ($n = 4$ cells). Stiffness was measured every minute during the 5–10 min periods when the cell was treated with each EDTA concentration. Average-force curves were registered with respect to the onset of force increase as the cantilever was lowered. **b**, Stiffness of a representative cell versus time, as measured with AFM. At $t = 10$ min the cell was treated with 2 µg ml^{-1} FM 4-64 ($n = 3$ cells). **c**, The ratio of the cell stiffness computed with AFM (Fig. 3c) versus the ratio of envelope stiffness computed from plasmolysis–lysis experiments (Fig. 2g) across chemical and genetic perturbations. The solid line is the linear best fit (linear regression: $R^2 = 0.27$, $F = 1.4$, not significantly different from horizontal). The numbers of cells used for each measurement are the same as given for Fig. 3c (AFM measurements of cell stiffness) and Fig. 2d–f (envelope stiffness). For the O8-expressing strain, we conservatively used a stiffness ratio of 1.5 for the fit. **d**, Calibration of AFM measurements using a polydimethylsiloxane sample with known Young's modulus of 3.5 MPa. Top, distribution of measurements of Young's modulus across the calibration sample. Red curve is the Gaussian fit to the data. Dashed line is the mean. Top inset, Young's modulus measurements were spatially uniform. Bottom, box plot of the distribution of Young's modulus measurements showing the median stiffness (red line), 25% and 75% percentiles (edges of box), extreme bounds (whiskers), and outliers (red points).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 7 | Genetic perturbations to the outer membrane render cells vulnerable to mechanical perturbation. a, Cell lysis versus time and cycle number during application of 5% detergent (*N*-lauroyl sarcosine, $n = 112$ cells) and 5% detergent with 400-mM oscillatory osmotic shock ($n = 100$ cells). Cell-lysis rate increased notably during 400-mM oscillatory osmotic shocks. **b**, Cell lysis versus time and cycle number during 400-mM oscillatory osmotic shocks with 2-min period for mutant *E. coli* strains with the MC4100 wild-type background ($n = 476$, 98, 187, 99, 280 cells for the wild-type, *imp4213*, Δ*ompA*, Δ*lpp* and Δ*pal* groups, respectively). **c**, Cell lysis versus time and cycle number during 400-mM oscillatory osmotic shock with 2-min period for *E. coli* wild-type AB1133 ($n = 273$ cells) and ATM378 (AB1133+O8, $n = 123$ cells). **d**, Time at which 25% of cells had lysed (ratio to the untreated wild type, Fig. 4b) versus the ratio of amplitudes during 100-mM oscillatory osmotic shocks (Fig. 2j). The line is the linear best fit (linear regression; $R^2 = 0.84$, $F = 19.4$, $P = 0.0045$). **e**, Serial dilutions of overnight *E. coli* L-form cultures spotted onto solid media permissive for L-form growth. For chemical treatments, 10 mM EDTA and 2 μg ml^{-1} FM 4-64 were included in the liquid media used to culture L-forms overnight, but not in the solid media onto which L-forms were spotted. Mutants (Δ*ompA*, +O8, Δ*lpp*, Δ*pal*) formed very small colonies, and were thus viewed using an inverted microscope to ensure accurate counting.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0346-1

Single-cell mapping of the thymic stroma identifies IL-25-producing tuft epithelial cells

Chamutal Bornstein[1,6], Shir Nevo[1,6], Amir Giladi[1,6], Noam Kadouri[1,6], Marie Pouzolles[2], François Gerbe[3], Eyal David[1], Alice Machado[2], Anna Chuprin[1], Beáta Tóth[4], Ori Goldberg[5], Shalev Itzkovitz[4], Naomi Taylor[2], Philippe Jay[3], Valérie S. Zimmermann[2], Jakub Abramson[1,7]* & Ido Amit[1,7]*

T cell development and selection are coordinated in the thymus by a specialized niche of diverse stromal populations[1–3]. Although much progress has been made over the years in identifying the functions of the different cell types of the thymic stromal compartment, there is no comprehensive characterization of their diversity and heterogeneity. Here we combined massively parallel single-cell RNA-sequencing[4,5], spatial mapping, chromatin profiling and gene targeting to characterize de novo the entire stromal compartment of the mouse thymus. We identified dozens of cell states, with thymic epithelial cells (TECs) showing the highest degree of heterogeneity. Our analysis highlights four major medullary TEC (mTEC I–IV) populations, with distinct molecular functions, epigenetic landscapes and lineage regulators. Specifically, mTEC IV constitutes a new and highly divergent TEC lineage with molecular characteristics of the gut chemosensory epithelial tuft cells. Mice deficient in *Pou2f3*, a master regulator of tuft cells, have complete and specific depletion of mTEC IV cells, which results in increased levels of thymus-resident type-2 innate lymphoid cells. Overall, our study provides a comprehensive characterization of the thymic stroma and identifies a new tuft-like TEC population, which is critical for shaping the immune niche in the thymus.

The thymus constitutes a specialized lymphoid organ, where immature T lymphocytes are educated to recognize foreign antigens, while tolerating self[1]. The T cell 'educational program' involves two central steps, which occur in two anatomical compartments of the thymus, the cortex and the medulla. Both compartments are characterized by the presence of specialized stromal cells, which provide the desired microenvironment for different checkpoints in T cell development and selection[2,3]. Cortical thymic epithelial cells (cTEC) coordinate the early stages of T cell development and positive selection of thymocytes[6]. The later steps of T cell development, including negative selection, are primarily carried out by mTECs[1–3].

Although much progress has been made over the years in elucidating the function of the different cell types of the thymic stroma, their diversity, heterogeneity and molecular pathways are still poorly characterized. To de novo characterize the entire stromal compartment of the thymus, we performed massively parallel single-cell RNA-sequencing (MARS-seq)[4,5] of 1,825 non-haematopoietic cells ($CD45^-$) isolated from adult mouse thymi (Extended Data Fig. 1). In order to link the canonical surface markers to the single-cell RNA-sequencing data, we used an index sorting strategy that allowed for retrospective analysis of surface markers of each individual cell. We then used the MetaCell pipeline to identify homogeneous and robust groups of cells (Methods). This analysis showed that the thymic stroma is composed of three major lineages, consisting of fibroblasts (*Col1a1* and *Col6a1*), endothelial cells (*Pecam1* and *Flt1*) and epithelial cells (*Epcam* and various keratin genes; Fig. 1a–c and Extended Data Fig. 2). Of the three linages, the epithelial cells displayed the largest heterogeneity in gene-expression programs (Extended Data Fig. 2 and Supplementary Tables 1, 2), suggesting that they are more complex and heterogeneous than previously anticipated.

In order to comprehensively characterize the TEC compartment, we sorted additional 1,716 $CD45^-EpCAM^+$ single cells (Extended Data Figs 1, 3). Clustering analysis combined with two-dimensional projection of the epithelial cells from both datasets revealed dozens of different TEC subpopulations (Fig. 1d, e, Extended Data Fig. 3 and Supplementary Table 2). The TEC subpopulations clustered within five major molecular types, each distributed at a distinct position within the two-dimensional projection (Fig. 1d and Extended Data Fig. 3). Index sorting analysis using the canonical cortical (Ly51) and medullary (UEA1) markers revealed that only one of the groups corresponds to the $Ly51^+UEA1^-$ population and expresses cTEC-specific genes, including *Prss16*, *Psmb11* and *Ctsl*[2,3]. By contrast, the other four TEC populations stained positively for UEA1 and had no or low expression of Ly51 (Extended Data Fig. 3), suggesting that these cells reside in the medulla. This was further validated by immunofluorescence staining and single-molecule RNA fluorescence in situ hybridization using a panel of markers specific to the individual TEC subpopulations (Fig. 1f and Extended Data Fig. 3). Therefore, on the basis of these data, we reclassified the mTEC compartment into four major groups (mTEC I–IV), reflecting their distinct transcriptional and molecular characteristics. Specifically, mTEC I is characterized by high expression of *Itga6* and *Sca1* (also known as *Ly6a*) (Fig. 1e and Extended Data Fig. 3), expression of which have previously been associated with putative TEC progenitors[7]. The mTEC II population is characterized by specific expression of the canonical markers of mature mTECs, including high expression of *Aire*, *Fezf2*, *Cd40*, *H2-Aa* or *Cd74* (Fig. 1e and Extended Data Fig. 3). mTEC III represents a heterogeneous population expressing several unique genes (*Pigr*, *Ly6d*, *Spink5*, *Ivl* and *Krt10*), some of which have been linked to a putative population of mTEC that previously expressed AIRE (post-AIRE cells)[8,9] (Fig. 1e and Extended Data Fig. 3). Notably, the mTEC IV population does not express any classical mTEC or cTEC markers, but rather a unique set of genes such as *Lrmp*, *Avil*, *Trpm5*, *Dclk1*, *Gng13*, *L1cam* and *Sox9* (Fig. 1e and Extended Data Fig. 3).

To investigate the dynamics of the different TEC populations during early development, we performed MARS-seq analysis of 3,074 sorted $CD45^-EpCAM^+$ single cells isolated from thymi at major developmental stages: embryonic day 14.5 (E14.5), E18.5 and day 6 after birth (Extended Data Figs. 1, 4). Because the developing thymus may have additional cell types or states that are not observed in the adult, we associated embryonic TEC metacells with adult phenotypes only if a large fraction of their cell neighbours was of adult origin (Methods, Fig. 1g, Extended Data Fig. 4 and Supplementary Table 3). This analysis highlighted the dynamic changes in the TEC compartment during thymus organogenesis. Although most of the E14.5 TECs were relatively homogenous and expressed a large number of cTEC-specific

[1]Department of Immunology, Weizmann Institute of Science, Rehovot, Israel. [2]Institut de Génétique Moléculaire de Montpellier, University of Montpellier, CNRS, Montpellier, France. [3]IGF, CNRS, INSERM, University of Montpellier, Montpellier, France. [4]Department of Cell Biology, Weizmann Institute of Science, Rehovot, Israel. [5]Department of Pediatrics, Schneider Children's Medical Center, Petach Tikva, Israel. [6]These authors contributed equally: Chamutal Bornstein, Shir Nevo, Amir Giladi, Noam Kadouri. [7]These authors jointly supervised this work: Jakub Abramson, Ido Amit.
*e-mail: jakub.abramson@weizmann.ac.il; ido.amit@weizmann.ac.il

© 2018 Springer Nature Limited. All rights reserved.

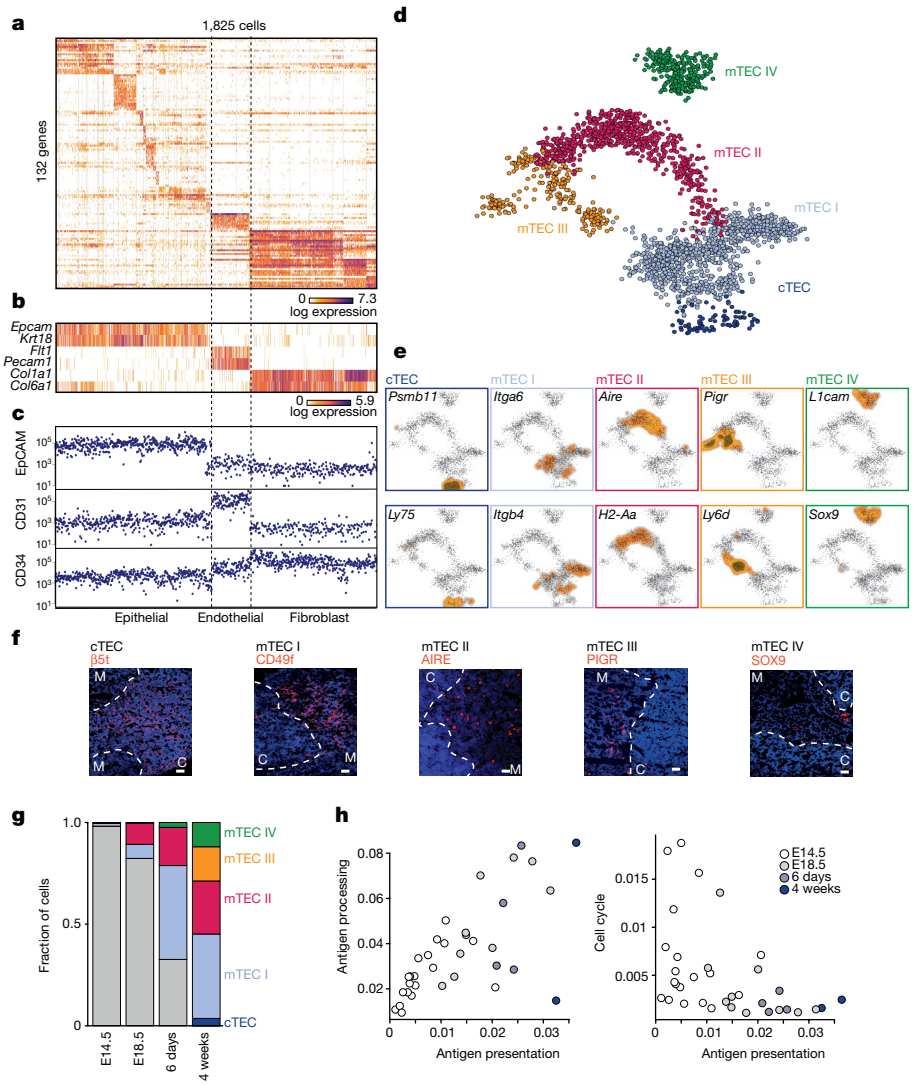


Fig. 1 | The medulla epithelial compartment has diverse molecular functions. a, Clustering analysis of 1,825 thymic stromal (CD45$^-$) cells from seven biological replicates of 4–6-week-old mice. **b**, Expression of selected marker genes. **c**, Index sorting tracks showing protein level intensity. **d**, Two-dimensional graphical representation of 2,341 single CD45$^-$EpCAM$^+$ cells separated into five TEC subsets. **e**, Kernel density projection of differentially expressed genes onto the two-dimensional graph. **f**, Immunofluorescence images of thymus sections. Medulla (M) and cortex (C) are separated by a dashed line, distinguished by nuclei density. Blue, DAPI. β5t is also known as PSMB11; CD49f is also known as ITGA6. Scale bars, 40 μm. Images are representative of three independent animals with similar results. **g**, Distribution of the TEC subsets along four developmental time points. Grey labels represent cells distinct from adult TECs. **h**, Progression of early cTECs towards adult cTECs. Circles represent cTEC metacells coloured by the developmental time points of the majority of cells. Axes represent share of each gene module from the entire metacell transcriptome.

genes, their general transcriptional signature was distinct from adult cTECs (Fig. 1g and Extended Data Fig. 4). Specifically, we observed progressive downregulation of cell cycle genes and upregulation of the MHC-II pathway in the adult compared to the embryonic cTECs (Fig. 1h, Extended Data Fig. 4 and Methods). While none of the adult mTEC subpopulations were present at the E14.5 stage, mTEC I and II, but not mTEC III and IV, became detectable in the thymus at E18.5 (Fig. 1g and Extended Data Fig. 4). At neonatal day 6, mTEC I, II and IV were present, although with different frequencies than in the adult thymus (Fig. 1g and Extended Data Fig. 4). Measuring the percentage of proliferating cells, we found that E14.5 TECs are the most proliferative cells in the thymus, while after birth most of the TEC proliferation is restricted to mTEC I and mTEC II cells (Extended Data Fig. 4).

To further characterize the newly defined mTEC subtypes, we established a new sorting strategy, based on a panel of surface markers unique to each population (Fig. 2a). This strategy was validated by MARS-seq and qPCR analyses of sorted EpCAM$^+$ populations, which were gated according to this scheme (Fig. 2b and Extended Data Fig. 5). Profiling the putative enhancer regulatory elements (marked by H3K4me2) of the four mTEC populations using indexing-first chromatin immunoprecipitation followed by deep sequencing[10] revealed that each subset is characterized by a unique set of distal enhancer regions with the mTEC IV population showing the most distinct regulatory elements (Fig. 2c, d, Extended Data Fig. 5 and Extended Data Table 1). In order to determine specific transcription-factor-binding sites, we performed an assay of transposase-accessible chromatin and analysed peaks of open chromatin within enhancer regions[11], followed by de novo motif finding (Extended Data Fig. 5 and Extended Data Table 1). This analysis revealed that accessible enhancer regions in mTEC II cells are significantly enriched ($P = 10^{-109}$; binomial test) for a binding-motif signature of the NF-κB family, correlating with specific expression of the *Nfkb2* gene in mTEC II cells (Fig. 2d–f and Extended Data Fig. 5). By contrast, mTEC-IV-specific enhancers were significantly enriched ($P = 10^{-805}$) for the POU class 2 transcription-factor motif, correlating with the specific expression of the *Pou2f3* gene in mTEC IV cells (Fig. 2d–f and Extended Data Fig. 5).

© 2018 Springer Nature Limited. All rights reserved.

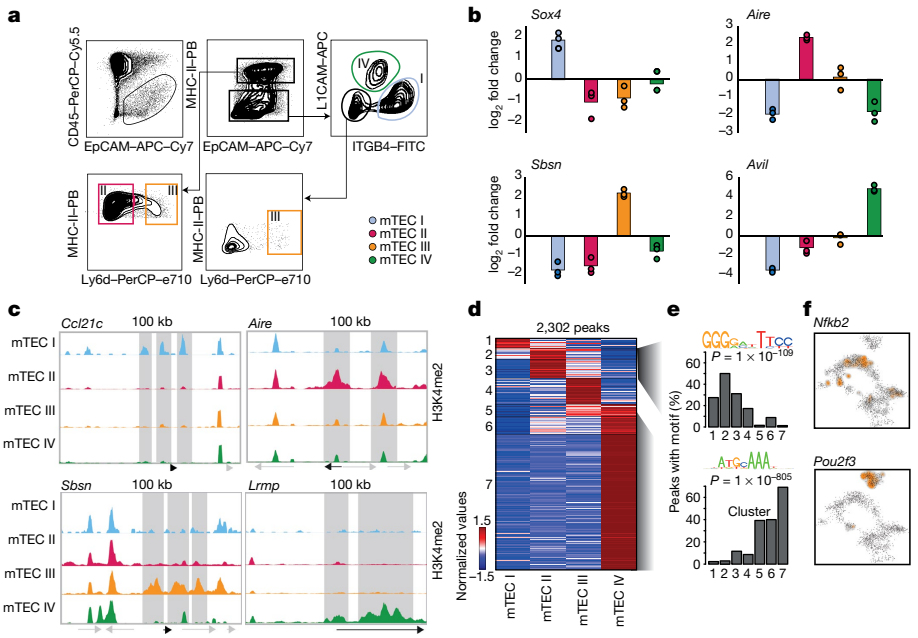


Fig. 2 | Genetic and epigenetic characterization of TEC subsets. a, Gating strategy for fluorescence-activated cell sorting to isolate mTEC subsets. **b**, qPCR analysis of mTEC subsets. Values represent fold change from the sample mean. $n = 3$ biologically independent animals. Data are mean (bars) and individual animals (dots). **c**, Normalized H3K4me2 profiles in 100-kb regions around differential mTEC gene loci. Differential genes are specified by black arrows. Plots are representative of four animals from two independent experiments. **d**, Peak intensities of 2,302 differential peaks between the four mTEC subsets (clusters 1–7; *K*-means; $K = 7$) **e**, Transcription-factor motif analysis showing enrichment of motifs in accessible regions within H3K4me2-marked peaks; bars indicate motif abundance in H3K4me2 peaks clusters 1–7. False-discovery rate-corrected binomial test; $n = 137$ and 1,559 peaks. **f**, Projection of transcription factors onto the two-dimensional graph (Fig. 1).

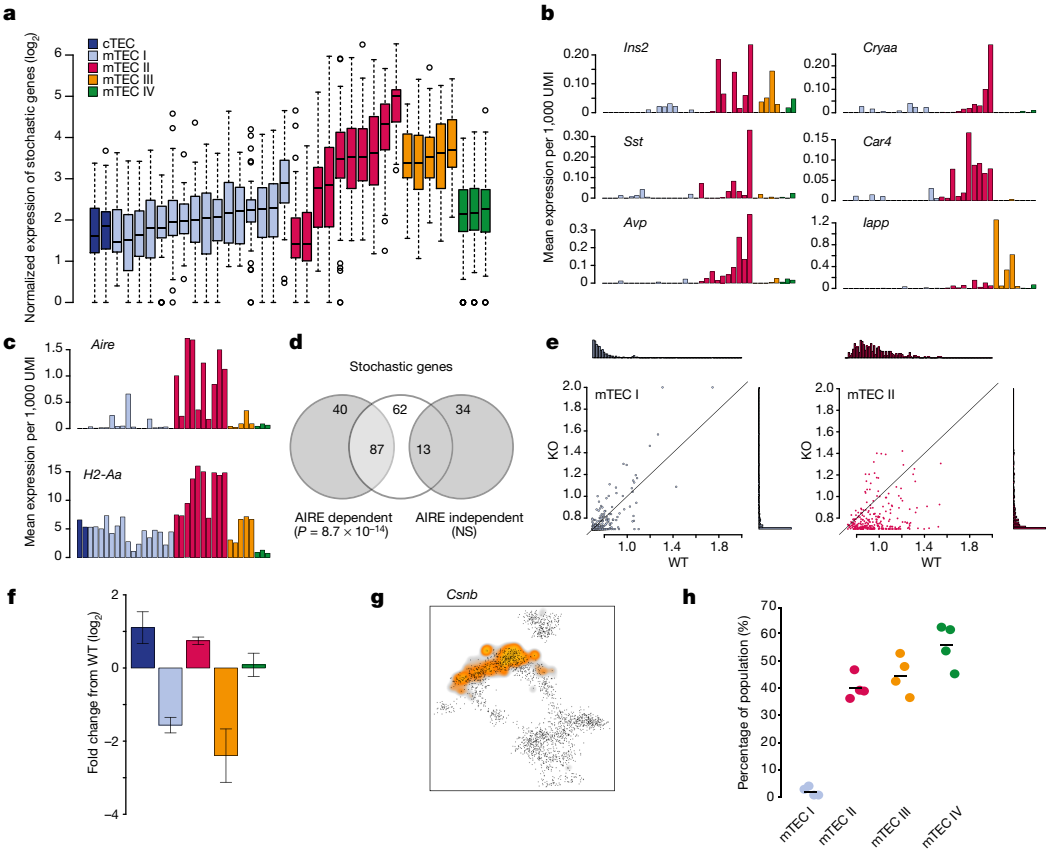


Fig. 3 | Characterization of AIRE-dependent mTEC subsets. a, Total expression distribution of stochastically expressed genes across the TEC metacells. In the box plots, bars indicate median, boxes are the first–third quantiles, whiskers, 5th–95th percentile and outlier are shown as circles. $n = 2{,}341$ single cells. **b**, **c**, Mean normalized expression of representative TRA genes (**b**), *Aire* and *H2-Aa* (**c**) across TEC metacells. **d**, Venn diagram depicting overlap between stochastically expressed genes and an established list of AIRE-dependent and -independent TRA genes. Hypergeometric test. $n = 388$ differentially expressed genes. **e**, Comparison of stochastic gene expression between *Aire* knockout (KO) and wild-type (WT) cells in TEC subsets. Marginal distributions are shown as histograms. Axes represent UMI count per 1,000 UMI, normalized to cell numbers. **f**, Bar plots showing $\log_2$ fold change between TEC subpopulation abundances in *Aire* knockout and wild-type mice. Error bars represent 95% confidence intervals. $n = 1{,}332$ knockout and 1,638 wild-type cells. **g**, Projection of *Csnb* onto the two-dimensional graph (Fig. 1). **h**, Percentage of tdTomato-expressing cells in mTEC I–IV. $n = 4$ biologically independent mice. Data are mean (line) and individual animals (dots).

© 2018 Springer Nature Limited. All rights reserved.

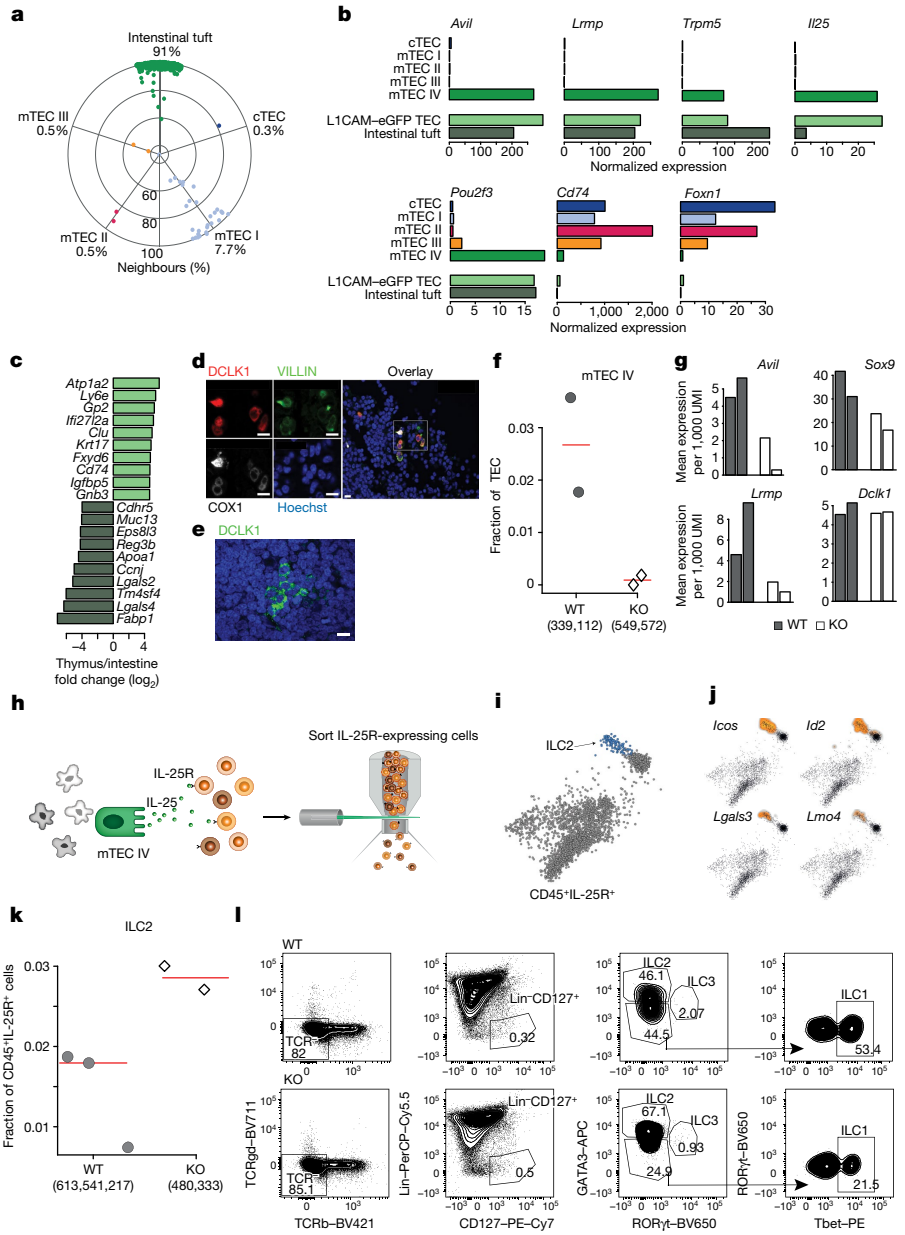


Fig. 4 | mTEC IV, a new TEC population with tuft-cell characteristics. a, *K*-nearest neighbours quantification ($K = 50$) of thymic L1CAM$^+$*Sox9*-eGFP$^+$ cells (mTEC IV) mapped to both TEC subsets and intestinal *Hpgds*-tdTomato$^+$ tuft cells (expressing tdTomato under *Hpgds* promoter). Radial location signifies populations with the highest number of neighbours; *y* axis indicates percentage of all nearest neighbours. **b**, Normalized mean expression of differentially expressed genes across TEC subsets, L1CAM$^+$*Sox9*-eGFP$^+$ ($n = 376$) and intestinal *Hpgds*-tdTomato$^+$ ($n = 1{,}879$ single cells) cells. **c**, Differential gene expression between L1CAM$^+$*Sox9*-eGFP$^+$ and intestinal *Hpgds*-tdTomato$^+$ cells ($\log_2$ fold change). **d**, **e**, Representative immunofluorescence imaging of tuft markers in thymic medulla sections from adult mice (**d**) and human tissue (**e**). Blue, DAPI. Scale bar = 10 μm. Images are representative of two independent experiments with similar results. **f**, Fraction of thymic mTEC IV cells from total TEC numbers in *Pou2f3* knockout and wild-type mice. Numbers in brackets indicate analysed cells in biological duplicates, indicated by individual circles and diamonds. Horizontal lines represent mean value. **g**, Mean expression of tuft markers in mTEC IV cells from wild-type and *Pou2f3* knockout mice. **h**, Experimental flow for exploring effect of mTEC IV cells on IL-25R$^+$ immune cells. **i**, Two-dimensional graphical representation of 3,500 CD45$^+$IL-25R$^+$ single cells separated into metacells. Blue dots label ILC2s. **j**, Projection of ILC2 markers onto the two-dimensional graph. **k**, Fraction of thymic ILC2 cluster from total CD45$^+$IL-25R$^+$ numbers in *Pou2f3* knockout and wild-type mice. Numbers in brackets indicate analysed cells in each replicate. Horizontal lines represent mean value and circles and diamonds indicate individual mice. **l**, ILC subtypes in *Pou2f3* knockout and wild-type thymi assessed by flow cytometry. Representative plots are shown ($n = 4$ biologically independent animals).

Given that one of the key functional roles of mTECs is to ectopically express and cross-present a plethora of tissue-restricted antigens (TRAs)[1,2], we analysed the level of TRA gene expression within the TEC compartment. Because the expression of most TRAs is stochastic and AIRE-dependent[12,13], we first defined a list of stochastically expressed genes in mTECs, based on their high expression variance and low correlation to other genes (Fig. 3, Extended Data Fig. 6, Supplementary Table 4 and Methods). As expected, mTEC II cells expressed the highest number of variable and uncorrelated genes (Fig. 3a, b and Extended Data Fig. 6). Notably, mTEC III cells showed a high level of stochastic gene expression (Fig. 3a–c) in spite of low expression of AIRE. MARS-seq analysis of 1,332 CD45$^-$EpCAM$^+$ single cells from *Aire* knockout mice, validated that AIRE deficiency almost entirely eliminated the expression of the 'stochastic' genes within the mTEC II and III populations (87% and 67%, respectively; Fig. 3e and Extended Data Fig. 6). In addition, AIRE deficiency also resulted in a decrease in mTEC I and mTEC III populations (Fig. 3f). By contrast, mTEC II cells and cTECs, showed an increase in frequency, while the mTEC IV population was unaffected (Fig. 3f). In order to better understand the lineage relationships of the individual mTEC subsets we performed in vivo fate mapping using *Csnb*^{cre}*Rosa26*^{tdTomato} reporter mice (Extended Data Fig. 6 and Methods). Whereas *Csnb* (also known as *Csn2*) expression was restricted to the mTEC II and III subsets (Fig. 3g and Extended Data Fig. 6), the tdTomato reporter was expressed in mTEC II, III and IV, but was absent from mTEC I cells (Fig. 3h and Extended Data Fig. 6). This suggests that while mTEC IV cells may be developmentally derived from the *Csnb*$^+$ mTEC II and/or III populations, or from a common ancestor, the mTEC I population is not.

Molecularly, the mTEC IV population is distinct from the other mTEC subsets, including the chromatin state, gene-expression profile and lack of stochastic gene expression (Figs. 1–3). On the basis of these data, we hypothesized that it may have a different functional role. This was further supported by specific expression of several genes that are associated with a rare epithelial lineage that is found in the gut, known as tuft (brush) cells[14–16] (Extended Data Fig. 7). In order to validate whether mTEC IV cells represent a putative tuft cell type, we compared their transcriptional profile to intestinal tuft cells and to the different TEC populations. Notably, mTEC IV cells were more similar to intestinal *Hpgds*-tdTomato$^+$ tuft cells than to any of the TEC subpopulations (Fig. 4a). Specifically, mTEC IV cells and intestinal tuft cells shared a large number of regulatory factors and tuft cell-specific genes including *Avil*, *Il25* and *Pou2f3*. However, they also showed differential gene

© 2018 Springer Nature Limited. All rights reserved.

expression, including for *Gp2* and *Gnb3*, which were only expressed in mTEC IV cells (Fig. 4b, c and Extended Data Fig. 7). Moreover, microscopy analysis of mTEC IV cells showed a typical tuft-like staining pattern in samples from both mice and humans (Fig. 4d, e). Finally, deficiency of a master regulator of intestinal tuft cells, *Pou2f3*, resulted in the complete loss of the mTEC IV population without affecting the development of any other TEC population (Fig. 4f, g and Extended Data Fig. 8), suggesting that they represent a bona fide tuft cell population.

Because the mTEC IV population is characterized by high and exclusive expression of IL-25 in the stroma of the thymus (Fig. 4b), we next studied their potential impact on thymic cells expressing the IL-25 receptor (*IL17rb*) (Fig. 4h). To this end, we first characterized 3,500 thymic $CD45^{+}IL\text{-}25R^{+}$ cells using MARS-seq and found five different subpopulations with distinct transcriptional states, including several $CD3^{+}$ thymocyte subsets, as well as thymus-resident type-2 innate lymphoid cells (ILC2s) (Fig. 4h–j and Extended Data Fig. 9). Notably, the loss of mTEC IV cells in *Pou2f3*$^{-/-}$ mice was accompanied by a significant increase in the thymic ILC2 population (Fig. 4k), whereas it had no significant impact on other thymic $CD45^{+}IL\text{-}25R^{+}$ cells or the main T cell subsets (Extended Data Fig. 9). The increase in the ILC2 ($Lin^{-}TCR^{-}CD127^{high}Tbet^{-}ROR\gamma t^{-}GATA3^{+}$) subset in *Pou2f3*$^{-/-}$ versus wild-type mice was further confirmed by conventional flow cytometry (Fig. 4l and Extended Data Fig. 9).

In summary, our study provides a comprehensive atlas of the stromal populations in the thymus of adult mice and during differentiation. We uncover unexpected complexity and diversity in this compartment, including an mTEC population with tuft-cell characteristics. This study also highlights important qualitative differences between the embryonic and postnatal thymus, which have, thus far, remained uncharacterized. As such, it clarifies many past controversies and confusions about the molecular, developmental and functional characteristics of the previously characterized TEC subsets and TRA expression, enabling the field to progress forward on a stable and common molecular blueprint. Nevertheless, several important open questions remain to be addressed in the future, including other functions of the mTEC IV population in the thymus and the lineage relationship and origins of the individual TEC subsets. Specifically, whether the mTEC II, III and IV populations are derived from a common ancestor cell, or whether mTEC IV cells are derived from the mTEC II and/or III populations.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0346-1.

Received: 5 September 2017; Accepted: 6 June 2018;
Published online 18 July 2018.

1. Klein, L., Kyewski, B., Allen, P. M. & Hogquist, K. A. Positive and negative selection of the T cell repertoire: what thymocytes see (and don't see). *Nat. Rev. Immunol.* **14**, 377–391 (2014).
2. Abramson, J. & Anderson, G. Thymic epithelial cells. *Annu. Rev. Immunol.* **35**, 85–118 (2017).
3. Takahama, Y., Ohigashi, I., Baik, S. & Anderson, G. Generation of diversity in thymic epithelial cells. *Nat. Rev. Immunol.* **17**, 295–305 (2017).
4. Jaitin, D. A. et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
5. Paul, F. et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell* **163**, 1663–1677 (2015).
6. Takada, K. & Takahama, Y. Positive-selection-inducing self-peptides displayed by cortical thymic epithelial cells. *Adv. Immunol.* **125**, 87–110 (2015).
7. Wong, K. et al. Multilineage potential and self-renewal define an epithelial progenitor cell population in the adult thymus. *Cell Rep.* **8**, 1198–1209 (2014).
8. Galliano, M. F. et al. Characterization and expression analysis of the *Spink5* gene, the mouse ortholog of the defective gene in Netherton syndrome. *Genomics* **85**, 483–492 (2005).
9. Hale, L. P. & Markert, M. L. Corticosteroids regulate epithelial cell differentiation and Hassall body formation in the human thymus. *J. Immunol.* **172**, 617–624 (2004).
10. Lara-Astiaso, D. et al. Chromatin state dynamics during blood formation. *Science* **345**, 943–949 (2014).
11. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
12. Brennecke, P. et al. Single-cell transcriptome analysis reveals coordinated ectopic gene-expression patterns in medullary thymic epithelial cells. *Nat. Immunol.* **16**, 933–941 (2015).
13. Meredith, M., Zemmour, D., Mathis, D. & Benoist, C. Aire controls gene expression in the thymic epithelium with ordered stochasticity. *Nat. Immunol.* **16**, 942–949 (2015).
14. Gerbe, F. et al. Distinct ATOH1 and Neurog3 requirements define tuft cells as a new secretory cell type in the intestinal epithelium. *J. Cell Biol.* **192**, 767–780 (2011).
15. Gerbe, F. et al. Intestinal epithelial tuft cells initiate type 2 mucosal immunity to helminth parasites. *Nature* **529**, 226–230 (2016).
16. Gerbe, F., Brulin, B., Makrini, L., Legraverend, C. & Jay, P. DCAMKL-1 expression identifies tuft cells rather than stem cells in the adult mouse intestinal epithelium. *Gastroenterology* **137**, 2179–2180 (2009).

Acknowledgements I.A. is supported by the Chan Zuckerberg Initiative, the HHMI International Scholar award, the European Research Council Consolidator Grant (ERC-COG)724471-HemTree2.0, the Israel Science Foundation (703/15), the Ernest and Bonnie Beutler Research Program of Excellence in Genomic Medicine, the Helen and Martin Kimmel award for innovative investigation, a Minerva Stiftung research grant, the Israeli Ministry of Science, Technology, and Space, the David and Fela Shapell Family Foundation, the NeuroMac DFG/Transregional Collaborative Research Center Grant, and the Abramson Family Center for Young Scientists. I.A. is the incumbent of the Alan and Laraine Fischer Career Development Chair. J.A. is supported by the ERC-2016-CoG-724821, Israel Science Foundation (1796/16 and 722/14), the Sy Syms Foundation; US–Israel Binational Foundation, Maurice and Vivienne Wohl Charitable Foundation; Goodman Family Charitable Lead Annuity Trust; Ruth and Samuel David Gameroff Family Foundation. J.A. is an incumbent of the Dr. Celia Zwillenberg-Fridman and Dr. Lutz Fridman Career Development Chair; A.G. is a recipient of the Clore fellowship; N.T., P.J. and V.Z. are supported by ANR-17-CE15-TUFTEFF, the Labex EpiGenMed and SIRIC Montpellier Cancer Grant INCa_Inserm_DGOS_12553; M.P. is supported by the Labex EpiGenMed and the FRM; F.G. and V.Z. are supported by CNRS; and N.T. and P.J. are supported by Inserm. We thank M. Boyer-Clavel and S. Gailhac of Montpellier Rio Imaging for cell sorting and the RAM animal facility of the IGMM.

Author contributions C.B., S.N., A.G., N.K., J.A. and I.A. designed the project, planned the experiments and wrote the manuscript. C.B., S.N. and N.K. performed experiments, A.G. and E.D. analysed the data, M.P., F.G., A.M., N.T., P.J. and V.Z. performed intestinal tuft and *Pou2f3* knockout mouse experiments. A.C. generated the *Csnb*-reporter mouse. B.T. and S.I. contributed to the single-molecule RNA fluorescence in situ hybridization experiment. O.G. provided human samples.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0346-1.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0346-1.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.A. or I.A.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Mice. All mice were maintained under specific pathogen-free conditions at the Weizmann Institute's animal facility and were handled in accordance with the guidelines of the Institutional Animal Care and Use Committee (25720316-2). Wild-type C57BL/6 (B6) mice were purchased from Harlan Laboratories. *Aire* knockout C57BL/6 mice were purchased from Jackson Laboratories. *Sox9*^{eGFP} mice were generated as described previously[17]. *Pou2f3* knockout and *Hpgds*^{tdTomato} mice were maintained under pathogen-free conditions at the IGMM and IGF animal facilities (RAM). *Csnb*^{cre} mice were generated using bacterial artificial chromosome (BAC) recombination. In brief, homology arms corresponding to a sequence downstream to the ATG of the *Csnb* gene were cloned into a Cre-recombinase-coding plasmid (pConst-Cre; provided by G. Schutz), followed by homologous recombination of the *cre* coding sequence into the BAC containing the *Csnb* gene (RP23-218H23, CHORI). Flp recombination was then used to remove the antibiotic resistance cassette. The final BAC was injected into fertilized BALB/c oocytes by the Weizmann transgenic core facility. Heterozygous mutants were backcrossed to C57BL/6 mice. *Csnb*^{cre}*Rosa26*^{tdTomato} reporter mice were generated by breeding heterozygous *Csnb*^{cre} mice with *Rosa26*^{tdTomato} mice (Jackson Laboratories).

Human samples. Human thymus samples were obtained during the course of corrective cardiac surgery at Schneider Children's Medical Center of Israel, following ethical approval (0781-16-RMC). An informed consent was signed by parents of the patients before obtaining the thymus.

Ethical compliance. All animals were housed according to guidelines at the Weizmann Institute of Science and the IGMM and IGF animal facilities (RAM). All experimental procedures were approved by the Institutional Animal Care and Use Committee (IACUC), application number 25720316-2.

All human experiments were done according to guidelines at the Weizmann Institute of Science and approved by Helsinki ethics committee, approval number 0781-16-RMC.

Isolation of mouse thymic stromal cells. Thymi from 4–8-week-old mice (unless otherwise stated, for embryonic or neonatal thymi) were placed into cold 1× PBS supplemented with 2% fetal bovine serum (FBS, Invitrogen). Thymi were chopped into small pieces and disintegrated by enzymatic digestion for 30–40 min in a 37 °C water bath, using 0.3 mg ml^{-1} collagenase D (Roche, 1088858), 1 mg ml^{-1} dispase II (Roche, 04942078001) and 10 ng ml^{-1} DNase I (Sigma-Aldrich, DN25) in RPMI supplemented with 2% FBS. Cells were then filtered through a 50-μm mesh filter and washed with 5–10 ml MACS buffer (1× PBS with 5 mM EDTA and 2% FBS), followed by centrifugation at 230*g* for 4 min. Percoll gradient density centrifugation was performed in order to enrich the stromal compartment. In brief, cells were resuspended in 2 ml of 1.115 g ml^{-1} isotonic Percoll (Sigma-Aldrich, P1644) and placed at the bottom of a tube. Subsequently, 1 ml of isotonic 1.065 g ml^{-1} Percoll and then 1 ml of 1× PBS were layered on top. The Percoll gradient was centrifuged at 2,700 r.p.m., 4 °C, with no deceleration for 30 min. The thymic stroma accumulated between the top and middle layers, and was collected and washed with MACS buffer and centrifuged at 230*g* for 4 min. Embryonic thymi were not subjected to Percoll separation, but were treated with red blood cell lysis buffer (15 mM ammonium chloride, 1 mM potassium bicarbonate, 10 μM EDTA in double distilled water, pH 7.3).

Isolation of thymic haematopoietic cells. Thymi from 6–12-week-old mice were surgically removed and placed in PBS on ice. Thymi were trimmed of fat and connective tissues, and thymocytes were extracted into a single-cell suspension by pressing the thymic lobes against a 70-μm cell strainer. Cells were washed in MACS buffer (1× PBS with 2% FBS and 5 mM EDTA pH 8.0).

Isolation of intestinal epithelial cells. Mouse small intestines were isolated, flushed with PBS and incised along their length. The tissue was incubated in 30 mM EDTA (Sigma-Aldrich) in HBSS pH 7.4 (Life Technologies) on ice, and transferred to DMEM (Life Technologies) supplemented with 10% FBS (Sigma-Aldrich). Vigorous shaking yielded the epithelial fraction that was then incubated with 100 μl of dispase (BD Biosciences) in 10 ml of HBSS, supplemented with 100 μl of DNase I at 2,000 Kunitz (Sigma). The single-cell preparation was obtained by filtration through a 30-μm mesh and used for further staining.

Antibodies and reagents for flow cytometry. APC–Cy7–EpCAM (118218), PE–Ly51 (108308), pacific blue–I-A/I-E (107620), PerCP–Cy5.5–CD45 (103132), FITC–CD34 (343603), APC–CD31 (102409), PE–Cy7–CD45 (103114), FITC–ITGB4 (123605), PE–ITGB4 (123610) and PE–IL-17RB (IL-25R,146305) were purchased from Biolegend; APC–L1CAM (FAB5674R) was purchased from Novus; APC–GATA3 (560078), brilliant violet 650–RORγt (564722), biotinylated CD3 (553060), biotinylated CD4 (553728), biotinylated CD8 (553029), biotinylated B220 (553086), biotinylated Ter119 (553672), biotinylated CD11c (553800), biotinylated Gr1 (553125), FITC–CD45 (553080), FITC–CD8 (553031), APC–Ter119 (557909), PE–Cy7–CD45 (552848), brilliant violet 711–CD4 (563726), APC–CD8 (553035), horizon V500–CD44 (560780), brilliant violet 711–Tcrgd (563994) and horizon V450–Tcrb (560706) were purchased from BD; and PerCP–eFluor710–Ly6d (46-5974-80), PE–Cy7–CD127 (25-1273-82), PE–Tbet (12-5825-82), eFluor780–CD4 (47-0042-82) and eFluor780–CD25 (47-025182) were purchased from eBioscience. The following materials were also used for FACS staining: 7-AAD Viability Staining Solution (Biolegend, 420404); Biotinylated Ulex Europaeus Agglutinin I (UEA I) (Vector laboratories, B-1065), PE–Cy7–streptavidin (405206, Biolegend), PerCP–Cy5.5–streptavidin (554064, BD), SYTOX (S34857, Invitrogen), FcR blocker (BE0307, Bio X Cell).

Flow cytometry and sorting. Cells were stained in MACS buffer (1× PBS with 2% FBS and 5 mM EDTA pH 8.0) with specific antibodies for 30 min at 4 °C. For haematopoietic cells, to avoid non-specific binding of antibodies to Fcγ receptors, anti-mouse CD16/CD32 monoclonal antibody (Bio X cell, BE0307 or Biolegend, 10130) was added to the antibody mix. Following staining, cells were washed and resuspended in MACS buffer. Secondary staining with streptavidin was performed in a similar manner. After the wash, cells intended for intracellular staining were fixed and permeabilized using the eBioscience fixation/permeabilization kit according to kit instructions, and stained with antibodies of intracellular markers for 2 h at 4 °C. Cells were sorted on a BD FACSAria Special Order Research Product (SORP) or BD FACSAria Fusion/II or III, or analysed on a BD LSRFortessa or BD FACSCantoII. Spectral overlap between fluorescent dyes was compensated using single-stained controls. Pre-gating was first done for live cells (in non-fixed samples) based on a 7-AAD, DAPI or SYTOX stain, followed by single-cell gating according to the FSC-A versus FCS-W plot. Data analysis was performed using FlowJo software (Tree Star Inc.).

mTEC subsets (Fig. 2a) were sorted from CD45^{-}EpCAM^{+} cells using the following sub-sorting gates: mTEC I: MHC-II^{low}ITGB4^{+}L1CAM^{-}; mTEC II: MHC-II^{+}Ly6d^{-}; mTEC III: ITGB4^{-}L1CAM^{-}Ly6d^{+}; mTEC IV: MHC-II^{low}L1CAM^{+}.

Single-cell index sorting. Isolated cells were single-cell sorted into 384-well cell capture plates containing 2 μl of lysis solution and barcoded poly(T) reverse-transcription primers for single-cell RNA-seq. Barcoded single-cell capture plates were prepared with a Bravo automated liquid handling platform (Agilent), as described previously[4]. To record marker levels of each single cell, the FACS Diva 7 'index sorting' function was activated during single-cell sorting.

Library preparation for single-cell RNA sequencing. Single-cell libraries were prepared as previously described[4]. In brief, 384-well plates, which contained lysis buffer and barcoded reverse-transcription poly-T primers, were immediately spun down and placed on dry ice. For barcoding and reverse transcription, a reverse-transcription reaction mix was added and the plates were placed in a PCR machine set to the appropriate program. All barcoded samples were pooled, followed by addition of exonuclease to remove excess RT–PCR primers. A purification step using SPRI beads that bind cDNA and RNA was performed after this step, as well as after each of the following steps. The pooled single-stranded cDNA was converted to a double-stranded DNA using a designated kit, in order to perform in vitro transcription of RNA molecules. The template DNA was then removed using DNase, and the generated RNA was fragmented and ligated to barcoded Illumina adapters. Reverse transcription of this ligation product was done using primers specific for the Illumina adapters, and libraries of the resulting cDNA were generated and enriched by 12–15 PCR cycles.

Low-level processing and filtering. All RNA-seq libraries (pooled at equimolar concentration) were sequenced using the Illumina NextSeq 500 at a median sequencing depth of 16,289 reads per single cell. Sequences were mapped to mouse genome (mm9), demultiplexed and filtered as previously described[4,18], extracting a set of unique molecular identifiers (UMI) that define distinct transcripts in single cells for further processing. We estimated the level of spurious UMIs in the data using statistics on empty MARS-seq wells (median noise 2.8%; Extended Data Fig. 2). Mapping of reads was done using HISAT (version 0.1.6)[19]; reads with multiple mapping positions were excluded. Reads were associated with genes if they were mapped to an exon, using the UCSC genome browser as reference. Exons of different genes that shared genomic position on the same strand were considered a single gene with a concatenated gene symbol. Cells with less than 500 UMIs were discarded from the analysis. After filtering, cells contained a median of 1,711 unique molecules per cell.

Data processing and clustering. The MetaCell pipeline was used to derive informative genes and compute cell-to-cell similarity, to compute KNN graph covers and derive distribution of RNA in cohesive groups of cells (or metacells), and to derive strongly separated clusters using bootstrap analysis and computation of graph covers on resampled data. The MetaCell package is described in detail in Supplementary Note 1. Default parameters were used unless otherwise stated.

Clustering for Figs. 1, 2 was done on a combined set of cells from two sources: (1) 1,972 CD45^{-} thymic cells and (2) 1,542 CD45^{-}EpCAM^{+} thymic cells (Extended Data Fig. 2). Clustering resulted in 49 clusters. Clusters with increased expression of *Hbb-b1*, *Trbc2* or *C1qb*, which are markers for red blood cells, T cells or macrophages, respectively (mean expression >10 times the median across clus-

© 2018 Springer Nature Limited. All rights reserved.

ters), were marked as contaminants and discarded from further analysis (Extended Data Fig. 3). We performed hierarchical clustering over the clusters structure and divided clusters into epithelial, endothelial and fibroblast groups by cutree.

Mapping cells to an existing cluster model. Given an existing reference single-cell dataset and cluster model, and a new set of single-cell profiles, we extracted for each new cell the K ($K=10$) reference cells with top Pearson correlation on transformed marker gene UMIs as described above. The distribution of cluster memberships over these K-neighbours was used to define the new cell reference cluster (by majority voting), and was applied for visualizing new cells by weighted average of the x and y coordinates of the clusters.

Clustering of development TEC and comparison to the existing model. Clustering of TEC during embryonic and postnatal development (Fig. 3) was done on a combined dataset of 1,343 E14.5, 895 E18.5 and 836 6 days postnatal epithelial cells. Two-dimensional projection of the resulting clustering was produced by MetaCell. However, in order to maintain the structure of the two-dimensional projection in Fig. 2, while enabling the discovery of new transcriptional states, we computed the KNN structure of the combined developmental and mature TEC. Developmental clusters with more than 20% cells for which >20% of their neighbours are within the mature dataset were associated with the mature projection, and their two-dimensional coordinates were determined by their mature neighbours. All other developmental clusters were assigned their regular coordinates (Fig. 3a and Extended Data Fig. 6).

For further refinement of the early TEC population (as in Fig. 3g), all developmental metacells with mean expression >0.25 times the median across clusters were included.

Gene modules and cell cycle analysis. Identification of ribosomal, cell cycle or other broadly expressed gene modules was done by clustering genes in a downsampled UMI matrix (500 molecules per cell). We filtered genes with total molecule (UMI) count lower than 5 and variance-to-mean ratio lower than 1.2. Hierarchical clustering using Ward's method was performed for detecting 30 fine-grained clusters. After removing clusters with mean Pearson intracorrelation lower than 0.025, 26 gene modules were retained. Manual annotation of the gene clusters was performed. This resulted in the identification of 26 modules with 8–136 genes, among which the cell cycle module contained 73 genes. Ribosomal modules (86 genes) were excluded from clustering analysis. Expression of cell cycle genes is a good indicator of proliferation[20]. To determine proliferation status of genes, we examined the pooled normalized expression of the cell cycle module across genes. This measurement showed a bimodal distribution, correlating with the total UMI count of cells (Extended Data Fig. 4e).

GO enrichment. Gene enrichment analysis was done using metascape software with mouse whole-genome backup.

Analysis of stochastically expressed genes. In order to define stochastically expressed genes, genes with less than 50 total UMIs were discarded (list A). Of this list, genes for which the corrected variance was greater than 1 (by a linear fit of the variance to the mean expression) were defined as variably expressed (list B). Gene-to-gene Pearson correlations were then computed on the UMI values of genes from list B (normalized to cell size). In order to discard tightly correlated gene expression programs, only genes for which the third highest correlation with other genes was less than 0.25 were considered as stochastically expressed (list C; Supplementary Table 4).

An established list of TRA genes was taken from a previous publication[21]. AIRE dependency of these TRA was determined by a twofold reduction in *Aire* knockout mice measured in the previous study[21]. Both AIRE-dependent and -independent lists were intersected with list B for further analysis.

Immunofluorescence staining of frozen thymic sections. Thymi from 4–8-week-old female mice were embedded in OCT compound (Tissue-Tek, Sakura) and frozen on dry ice. Cryostat sections (6 μm) were fixed with ice-cold acetone for 10 min and incubated with primary antibody (anti-AIRE AF488-conjugated, 04-150; Millipore) diluted in 1% BSA in PBS for 60 min at room temperature. Sections were washed three times with PBS and incubated with DAPI staining for 10 min at room temperature, followed by three washes with PBS.

Immunofluorescence staining of PFA-fixed frozen thymic sections. Thymi from 4–8-week-old wild-type, $Sox9^{eGFP}$, $Csnb^{cre+}Rosa26^{tdTomato}$ and $Csnb^{cre-}Rosa26^{tdTomato}$ mice were isolated, cleaned and fixed for 3 h with ice-cold 3.7% formaldehyde, followed by overnight incubation with cryoprotection solution (3.7% formaldehyde, 30% sucrose in 1× PBS). Thymi were embedded in OCT and frozen on dry ice. Cryostat sections (6–7 μm) were permeabilized and blocked with blocking buffer (TBS, pH 7.4, 5% goat serum or BSA, and 0.1% Triton X-100) for 30 min at room temperature. Sections were then incubated with primary antibodies diluted in blocking buffer overnight at 4 °C. Primary antibodies used were as follows: anti-PSMB11 (pd021, MBL); anti-PIGR (AF2800, R&D Systems); anti-DCLK1 (ab37994, Abcam); APC-conjugated anti-ITGA6 (313615; Biolegend). Sections were washed twice with TBST (1× TBS supplemented with 0.1% Tween-20). Sections stained with unconjugated antibodies were incubated with a secondary antibody (goat anti-rabbit AF555, Jackson Laboratories) for 60 min at room temperature. DAPI staining was performed for 10 min at room temperature.

Immunoflourescence staining of paraffin-embedded thymic sections. Thymi from 4–8-week-old mice or 8-day-old male humans were fixed in 4% PFA for 48 h, followed by embedding in paraffin. Subsequently, 5-μm-thick sections were dewaxed in xylene and rehydrated in graded alcohol baths. Antigen retrieval was performed by boiling slides for 20 min in 10 mM sodium citrate buffer, pH 6.0. Nonspecific binding sites were blocked in blocking buffer (TBS pH 7.4, 5% goat serum and 0.1% Triton X-100) for 30 min at room temperature. Sections were incubated with primary antibodies diluted in blocking buffer overnight at 4 °C. Primary antibodies used were as follows: anti-VILLIN (MAB1671, Millipore); anti-COX1 (sc-1754, Santa Cruz Biotechnology); anti-DCLK1 (ab37994, Abcam). Slides were washed two times with TBST before incubation with fluorescent secondary antibodies conjugated to AF488, AF555, Cy3 or Cy5 (Jackson ImmunoResearch Laboratories) and Hoechst (Sigma-Aldrich) or DAPI in TBS–0.1% Triton X-100.

Single-molecule RNA fluorescence in situ hybridization (FISH). Single-molecule FISH probe libraries consisting of 48 probes with a length of 20 bp were designed as previously described[22], constructed and provided by Agentek. The following probe libraries were used: *Avil*, *Sbsn* coupled to Cy5, *Epcam* and *Pigr* coupled to AF594. RNA FISH was performed as previously described. In brief, thymi from 4–8-week-old female mice were isolated, cleaned and fixed for 3 h with ice-cold 3.7% formaldehyde followed by overnight incubation with cryoprotection solution (3.7% formaldehyde, 30% sucrose in PBS). Thymi were embedded in OCT and frozen on dry ice. Cryostat sections (5–8 μm) were air-dried and fixed again with 3.7% formaldehyde for 5 min following by 2 h incubation with 70% ethanol at 4 °C. Sections were rehydrated with 2× SSC, and treated with proteinase K in 2× SSC for 10 min at room temperature. Hybridization was performed overnight by 20% formamide in 2× SSC with 0.1 ng μl^{-1} of the desired probes at 30 °C. DAPI (to stain the nuclei) was added during the washes, and sections were incubated with fresh GLOX buffer (10 mM Tris pH 8.0 and 0.4% glucose in 2× SSC).

Imaging and image analysis. Imaging was performed on Ultima Multiphoton Microscope, Nikon Eclipse TI-S fluorescence microscope, or Nikon-Ti-E inverted fluorescence microscope with 60× and 100× oil-immersion objectives and a Photometrics Pixis 1024 CCD camera, using MetaMorph software (Molecular Devices) or NIS element software (Nikon). Image analysis was performed with ImageJ software.

Indexing first chromatin immunoprecipitation followed by deep sequencing (iChIP–seq). iChIP–seq was prepared as previously described[10]. In brief, cells were cross-linked for 8 min in 1% formaldehyde and quenched for 5 min in 0.125 M glycine before sorting. Cells were sorted using the described sorting strategy and frozen. Cell pellets were lysed in 0.5% SDS and sheared with the NGS Bioruptor Sonicator (Diagenode). Sheared chromatin was immobilized on 15 μl Dynabeads Protein G (Invitrogen) with 1.3 μg of anti-H3 antibody (Abcam). Magnetized chromatin was then washed with 10 mM Tris-HCl supplemented with 1× PI. Chromatin was end-repaired, dA-tailed and ligated with sequencing adapters containing Illumina P5 and P7 sequences. Indexed chromatin was pooled and incubated with 2.5 μg H3K4me2 antibody (ab32356, Abcam) at 4 °C for 3 h and for an additional hour with protein G magnetic beads (Invitrogen). Magnetized chromatin was washed and reverse cross-linked. DNA was subsequently purified with 1.65× SPRI and amplified by PCR with 0.5 μM of forward and reverse primers containing Illumina P5-rd1 and P7-rd2 sequences. Library concentration was measured with a Qubit fluorometer and mean molecule size was determined by TapeStation (Agilent). DNA libraries were sequenced on an Illumina NextSeq 500 with an average of over 10 million aligned reads per replicate.

ATAC-seq. To profile open chromatin, we used an assay of transposase-accessible chromatin following sequencing (ATAC-seq) as previously published[23], with modifications as previously described[10]. In brief, cell populations were sorted in 400 μl of MACS buffer (1× PBS, 0.5% BSA, 2 mM EDTA) and pelleted by centrifugation for 15 min at 500*g* and 4 °C with low acceleration and brake settings. Cell pellets were washed once with 1× PBS and cells were pelleted by centrifugation using the previous settings. Cell pellets were resuspended in 25 μl of lysis buffer (10 mM Tris-HCl (pH 7.4), 10 mM NaCl, 3 mM $MgCl_2$, 0.1% Igepal CA-630) and nuclei were pelleted by centrifugation for 30 min at 500*g*, 4 °C with low acceleration and brake settings. Supernatant was discarded and nuclei were resuspended in 25 μl reaction buffer containing 2 μl of Tn5 transposase and 12.5 μl of TD buffer (Nextera Sample preparation kit from Illumina). The reaction was incubated at 37 °C for 1 h. Then, 5 μl of clean-up buffer (900 mM NaCl, 300 mM EDTA), 2 μl of 5% SDS and 2 μl of Proteinase K (NEB) were added and incubated for 30 min at 4 °C. Tagmentated DNA was isolated using 2× SPRI beads clean-up. For library amplification, two sequential nine-cycle PCR runs were performed in order to enrich small tagmentated DNA fragments. We used 2 μl of indexing primers included in the Nextera Index kit and KAPA HiFi HotStart ready mix. After the first PCR, the libraries were selected for small fragments (less than 600 bp) using SPRI clean-up. Then a second PCR was performed with the same conditions in order to obtain the final library.

© 2018 Springer Nature Limited. All rights reserved.

Processing of iChIP–seq, ATAC-seq and chromatin peak calling. Reads were aligned to the mouse reference genome (mm9, NCBI v.37) using Bowtie aligner version 1.0.0[24] with best match parameters (bowtie -m 1–sam–best–strata -v 2). To identify regions of enrichment (peaks) from ChIP–seq reads of H3K4me2, we used the HOMER package makeTagDirectory followed by the findPeaks command with the histone parameter[25] and IDR filtering for reproducible peaks across replicates[26]. Union peaks file were generated by combining and merging overlapping peaks in all samples.

Chromatin analysis. For clustering of differential peaks, we first averaged peak sizes across replicates. We defined differential peaks as peaks for which the maximum value is more than fourfold higher than their minimum value. We then normalized peaks intensities and performed *K* means clustering ($K=7$). For motif finding, we independently called peaks in ATAC-seq, as above, and identified the maximum peak that overlapped each H3K4me2 region. The overlapping sequences were input for HOMER package motif finder algorithm findMotifGenome[25].

Gene tracks and visualization. All gene tracks were visualized as bigWig files of the combined replicates normalized to 10,000,000 reads, using Integrative Genomics Viewer (http://www.broadinstitute.org/igv).

Real-time PCR analysis. Cells were sorted into 40–50 μl of lysis/binding buffer (Life Technologies). mRNA was captured with 15 μl of Dynabeads oligo(dT) (Life Technologies), washed and eluted at 85 °C with 10 μl of 10 mM Tris-Cl (pH 7.5). The purified RNA was used for cDNA synthesis using the High-Capacity cDNA Reverse-Transcription kit (Applied Biosystems) and polyT primers. The subsequent qPCR analysis was performed using the Fast SYBR Green Master Mix (Life technologies). Differential expression was calculated according to the $\Delta\Delta C_t$ method. Specific qPCR primers: *Actb*: GGAGGGGGTTGAGGTGTT, TGTGCACTTTTATTGGTCTCAAG; *Sox4*: GCTGGGCTTTCTCCTCCT, AGGCTGGCCTGCTACTCC; *Aire*: TGGGCTGATTAGGACCAAGA, ACAAAGATCAGGGCCATCTG; *Avil*: GCATCAGGACCCACATCTGC, ATGCTGTGGCACATGGTAGAC; *Sbsn*: CACCATGCCCTAAACTGATGC, ACAAAGCTCAAAGCAGCCCTC; *cre*: AGGAGAATGTGGATGCTGGGG, CAATTTCGGCAATGCGCAGC; *Csnb*: AAACTTCAGAAGGTGAATCTCATGG, GCTGGATGTTTTGTGGGACG.

Lists of ribosomal and cell cycle modules. *Ribosomal genes. 2810422J05Rik*, *AC151602.1*, *AL663027.1*, *Cfl1*, *Eef1b2*, *Ftl1*, *Gm10059*, *Gm10076*, *Gm10443*, *Gm11361*, *Gm11808*, *Gm11942*, *Gm12630*, *Gm13408*, *Gm14456*, *Gm15427*, *Gm15459*, *Gm15710*, *Gm3788*, *Gm4149*, *Gm5244*, *Gm5559*, *Gm8730*, *Gm8759*, *Gnb2l1*, *Hspa8*, *Mif*, *Rpl10*, *Rpl13*, *Rpl13a*, *Rpl14*, *Rpl18a*, *Rpl22*, *Rpl26*, *Rpl28-ps3*, *Rpl29*, *Rpl32*, *Rpl34*, *Rpl37*, *Rpl37a*, *Rpl38*, *Rpl38-ps2*, *Rpl4*, *Rpl7a-ps12*, *Rpl8*, *Rpl9-ps6*, *Rplp0*, *Rplp1*, *Rplp2*, *Rps10*, *Rps10-ps1*, *Rps14*, *Rps15*, *Rps18*, *Rps19*, *Rps2*, *Rps20*, *Rps21*, *Rps26*, *Rps28*, *Rps3*, *Rps3a*, *Rps5*, *Rps8*, *Rps9*, *Rpsa*, *Rpsa-ps10*, *Snord35a*, *Tmsb10*, *Tpt1*, *B2m*, *Eef1a1*, *Gas5*, *Gm13456*, *Gm15500*, *Gm16247*, *Rpl23*, *Rpl35a*, *Rpl7*, *Rps11*, *Rps24*, *Rps25*, *S100a11*, *Tmsb4x*, *Ubl5*, *mmu-mir-703*. *Cell cycle genes. 2810417H13Rik*, *2900006K08Rik*, *4833427G06Rik*, *4930473A06Rik*, *5133401N09Rik*, *6820408C15Rik*, *Arhgap11a*, *Arl6ip1*, *Birc5*, *Ccdc108*, *Ccdc113*, *Ccdc151*, *Ccdc17*, *Ccdc19*, *Ccdc30*, *Ccdc39*, *Ccdc40*, *Ccdc67*, *Ccno*, *Cdc20*, *Cdca8*, *Cenpf*, *Ckap2*, *Ckap2l*, *Cks1b*, *Dcdc2a*, *Dek*, *Dnahc9*, *Dnajb13*, *E030019B06Rik*, *Foxm1*, *Gm11423*, *Gm9938*, *H2afx*, *Hist1h1a*, *Hist1h1b*, *Hist1h1d*, *Hist1h1e*, *Hist1h2ac*, *Hist1h2ae*, *Hist1h2ao*, *Hist1h2ap*, *Hist1h3c*, *Hist1h3e*, *Hist1h4d*, *Hmgb2*, *Hnrnpa2b1*, *Ift46*, *Kif15*, *Kif24*, *Lrrc23*, *Lrrc46*, *Mki67*, *Phospho2*, *Pih1d2*, *Plk1*, *Rsph1*, *Rsph9*, *Smc4*, *Supt16h*, *Tekt1*, *Tekt4*, *Tmem107*, *Top2a*, *Ttll6*, *Tuba1b*, *Tubb2c*, *Tubb5*, *Ube2c*, *Wdr52*, *Wdr65*, *Zmynd10*.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Code availability. No custom software was used to collect data. iChIP–seq and ATAC-seq data analyses, including motif finding, were done with the HOMER package[25]. Single-cell data were analysed with the MetaCell package, which is available upon request.

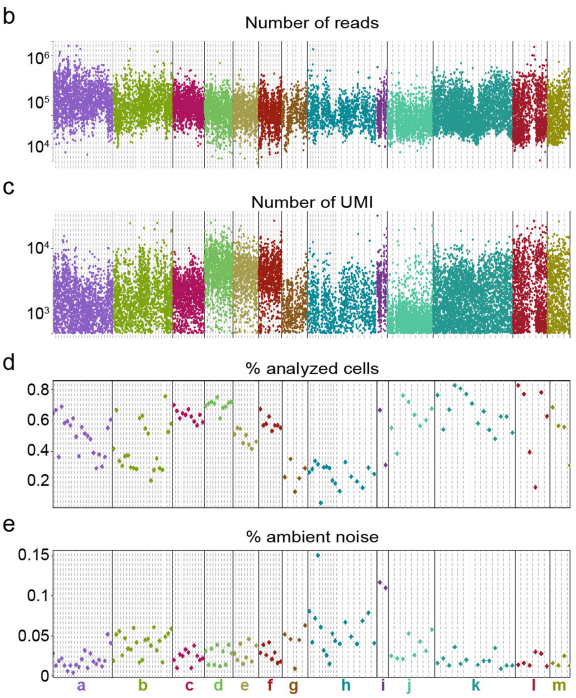
Data availability. RNA, iChIP–seq and ATAC-seq data reported in this paper were deposited with Gene Expression Omnibus under accession numbers: GSE103967, GSE103968, GSE103969 and GSE103970.

17. Tanimizu, N., Nishikawa, Y., Ichinohe, N., Akiyama, H. & Mitaka, T. Sry HMG box protein 9-positive ($Sox9^+$) epithelial cell adhesion molecule-negative ($EpCAM^-$) biphenotypic cells derived from hepatocytes are involved in mouse liver regeneration. *J. Biol. Chem.* **289**, 7589–7598 (2014).
18. Gury-BenAri, M. et al. The spectrum and regulatory landscape of intestinal innate lymphoid cells are shaped by the microbiome. *Cell* **166**, 1231–1246 (2016).
19. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
20. Scialdone, A. et al. Computational assignment of cell-cycle stage from single-cell transcriptome data. *Methods* **85**, 54–61 (2015).
21. Sansom, S. N. et al. Population and single-cell genomics reveal the *Aire* dependency, relief from Polycomb silencing, and distribution of self-antigen expression in thymic epithelia. *Genome Res.* **24**, 1918–1931 (2014).
22. Lyubimova, A. et al. Single-molecule mRNA detection and counting in mammalian tissue. *Nat. Protoc.* **8**, 1743–1758 (2013).
23. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: a method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **109**, 21.29.1–21.29.9 (2015).
24. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
25. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).
26. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5**, 1752–1779 (2011).

© 2018 Springer Nature Limited. All rights reserved.

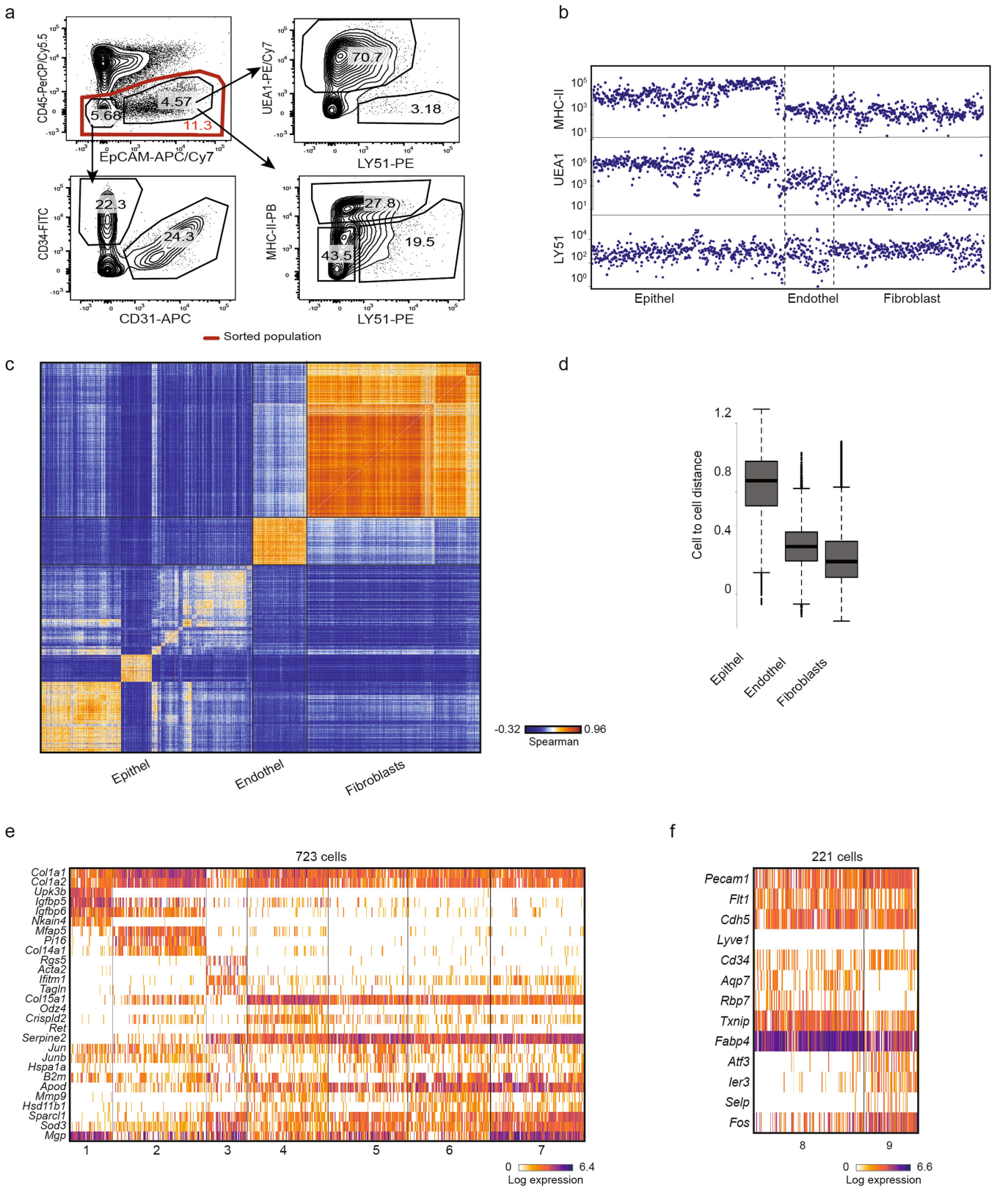
a

Organ	Gating	Age	Strain		Mouse	nbatches	ncells
Thymus	CD45⁻	4-6 weeks	WT	a	1	3	322
					2	3	358
					3	4	393
					4	2	206
					5	2	188
					6	5	332
					7	2	222
	CD45⁻ EpCAM⁺	4-6 weeks	WT	b	8	2	208
					9	4	267
					10	3	168
					11	4	443
					12	2	95
					13	3	178
					14	2	246
					15	1	111
			Aire KO	c	16	11	1332
		E14.5	WT	d	17	2	269
					18	2	274
					19	4	525
					20	2	275
		E18.4	WT	e	21	1	129
					22	2	220
					23	5	546
		6 days	WT	f	24	4	396
					25	4	440
		4-6 weeks	*Pou2f3* Control	g	26	4	339
					27	1	112
			Pou2f3 KO	h	28	11	549
					29	6	572
	CD45⁻ EpCAM⁺ *Sox9*-eGFP⁺	4-6 weeks	*Sox9*^{eGFP}	i	30	2	376
Intestine	EpCAM⁺ *Hpgds*-tdTomato⁺	4-6 weeks	*Hpgds*^{tdTomato}	j	31	4	928
					32	4	951
Thymus	CD45⁺ IL-25R⁺	4-6 weeks	WT	k	33	4	1075
					34	6	1560
					35	4	865
			Pou2f3 Control	l	36	2	613
					37	2	217
					38	2	541
			Pou2f3 KO	m	39	2	480
					40	2	333
					Total	135	17654



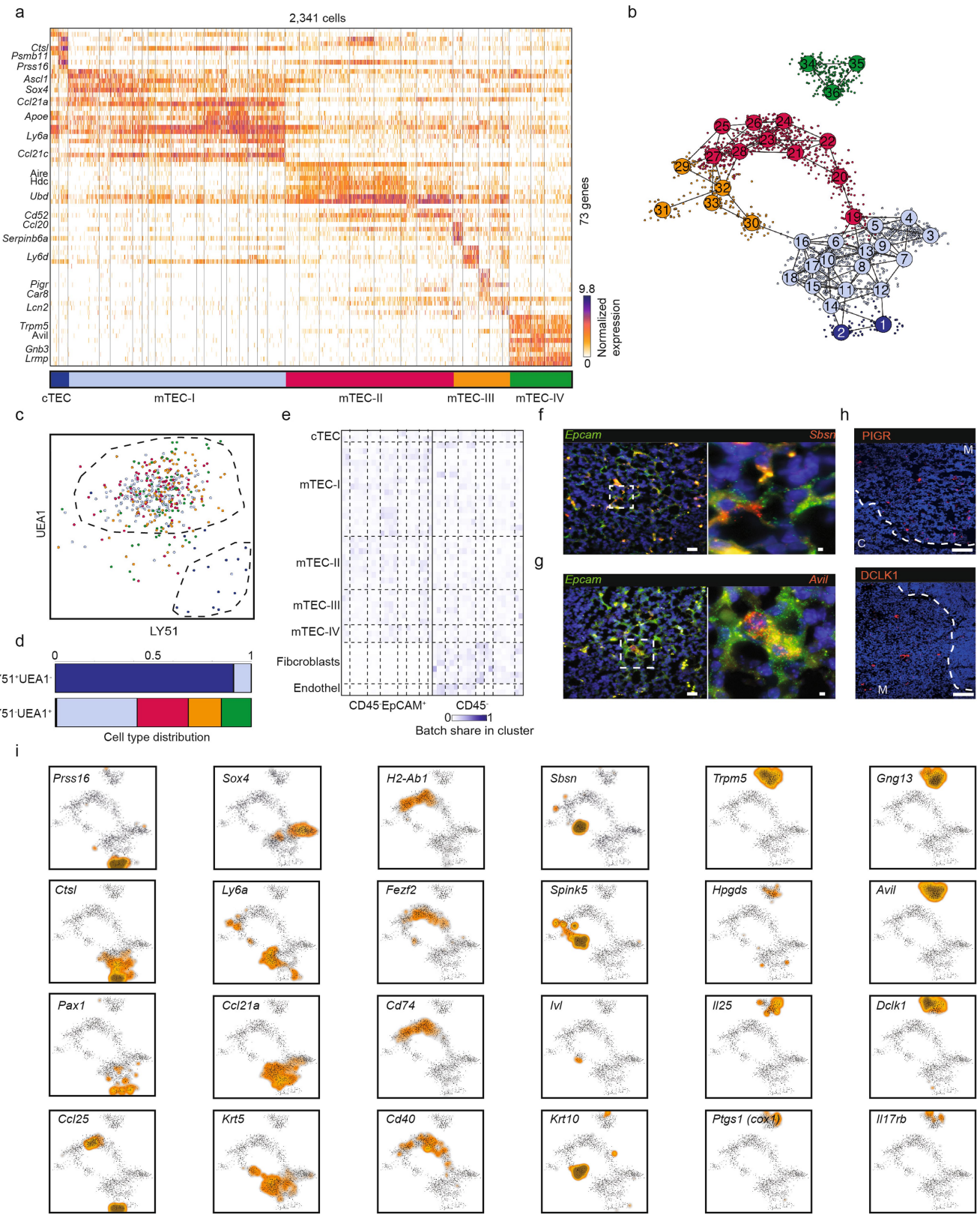
Extended Data Fig. 1 | Single-cell data quality controls. a, Summary of all single cells analysed in this study, divided into experimental procedures. 'nbatches' indicates number of technical replicates; 'ncells' indicates number of cells after filtering (see Methods). **b**–**e**, Colour-coded tracks summarizing the number of Illumina reads per cell (**b**), transcripts (UMI) detected in each cell (**c**), fraction of analysed cells from each amplification batch (**d**) and estimation of technical noise for each amplification batch (**e**). Cells are coloured by experimental procedure. Technical noise is assessed by genomic UMIs in empty wells as previously described[4] (see Methods).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 2 | Thymic stroma sorting and clustering. a, Flow cytometry schematic of thymic cells showing isolation of stroma cells, as well as staining for known populations markers. Immune cells, CD45; fibroblast, CD34; endothelial, CD31; mTEC, UEA1; cTEC, LY51; mature mTEC, MHC-II. The red border marks stroma single-cell sorting gate. **b**, Index sort tracks showing the intensity of protein levels for MHC-II, UEA1 and LY51 in individual single cells shown in Fig. 1a–c. **c**, Cell–cell correlation of CD45$^-$ thymic stroma cells calculated over 132 differentially expressed genes. **d**, Pairwise distance distribution between cells within the three main stromal lineages. Distance is defined as 1 − Spearman. Box plots display median bar, first–third quantile box and 5th–95th percentile whiskers. $n = 1{,}825$ single cells. **e**, **f**, Gene expression profiles of 723 cells from fibroblast clusters of the CD45$^-$ stroma, marked by increased expression of *Col1a1* (**e**, Supplementary Table 1) and 221 cells from the endothelial clusters of the CD45$^-$ stroma, marked by increased expression of *Pecam1* (**f**, Supplementary Table 1).

© 2018 Springer Nature Limited. All rights reserved.

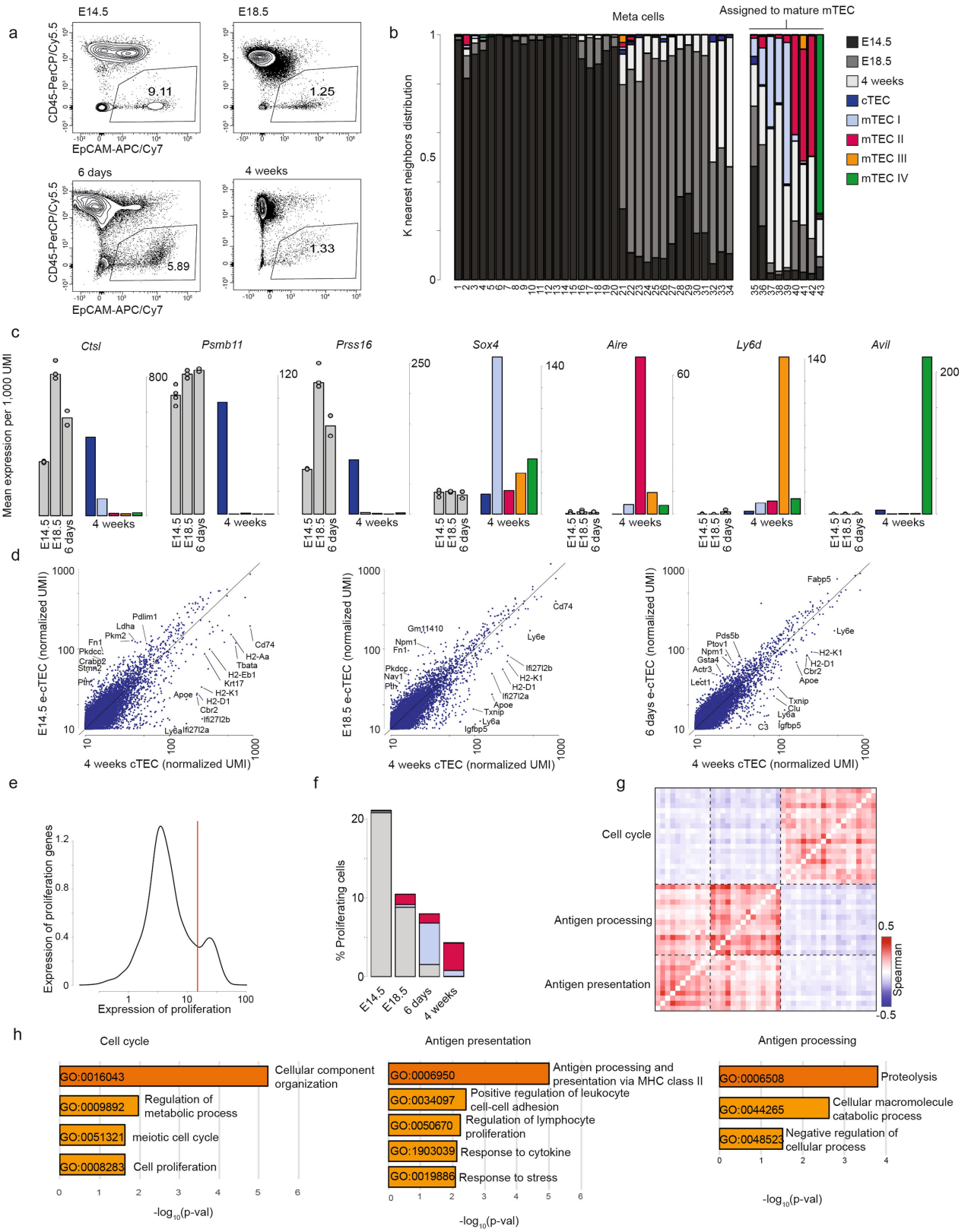


Extended Data Fig. 3 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 3 | Thymic epithelial cells are characterized by four subsets of mTEC and a single cTEC subset. a, Heat map showing a metacell analysis of 2,341 thymic epithelial cells ($CD45^{-}EpCAM^{+}$), featuring the 73 most variable genes, from 15 biological replicates of 4–6-week-old mice. Colour bar represents separation of 36 metacells into five main populations. **b**, Two-dimensional graph representation of the metacell model in Fig. 1a (see Methods). Big circles represent metacells, and are colour-coded as shown in **a**. **c**, FACS index sorting measurement of LY51 and UEA1 in epithelial cells. Cells are coloured based on cluster association as determined in **a**. Dashed lines outline $LY51^{+}UEA1^{-}$ and $LY51^{-}UEA1^{+}$ gates. **d**, Fraction of TEC subsets out of $LY51^{+}UEA1^{-}$ and $LY51^{-}UEA1^{+}$ populations, assessed by gating single cells on index sorting protein measurements of UEA1 and LY51 in **c**. **e**, Controlling for batch effect as determined by the relative share of each batch in all metacells. Batches are ordered by biological replicate (marked by dashed lines) and sorting scheme (either $CD45^{-}$ or $CD45^{-}EpCAM^{+}$). **f**, **g**, Single molecule FISH assay on 5–8-μm-thick cryosections (see Methods) using fluorescent probes against the genes *Epcam* and *Sbsn* (**f**) or *Avil* (**g**). Blue, DAPI. The experiments were repeated independently four times with similar results. **h**, Immunofluorescence images of the protein markers: PIGR and DCLK1. Medulla (M) and cortex (C) are separated by dashed lines, distinguished by nuclei density. Blue, DAPI. **f**–**h**, Scale bars, 20 μm. The experiments were repeated independently twice with similar results. **i**, Projection of representative differentially expressed genes onto the two-dimensional graph of epithelial cells.

© 2018 Springer Nature Limited. All rights reserved.

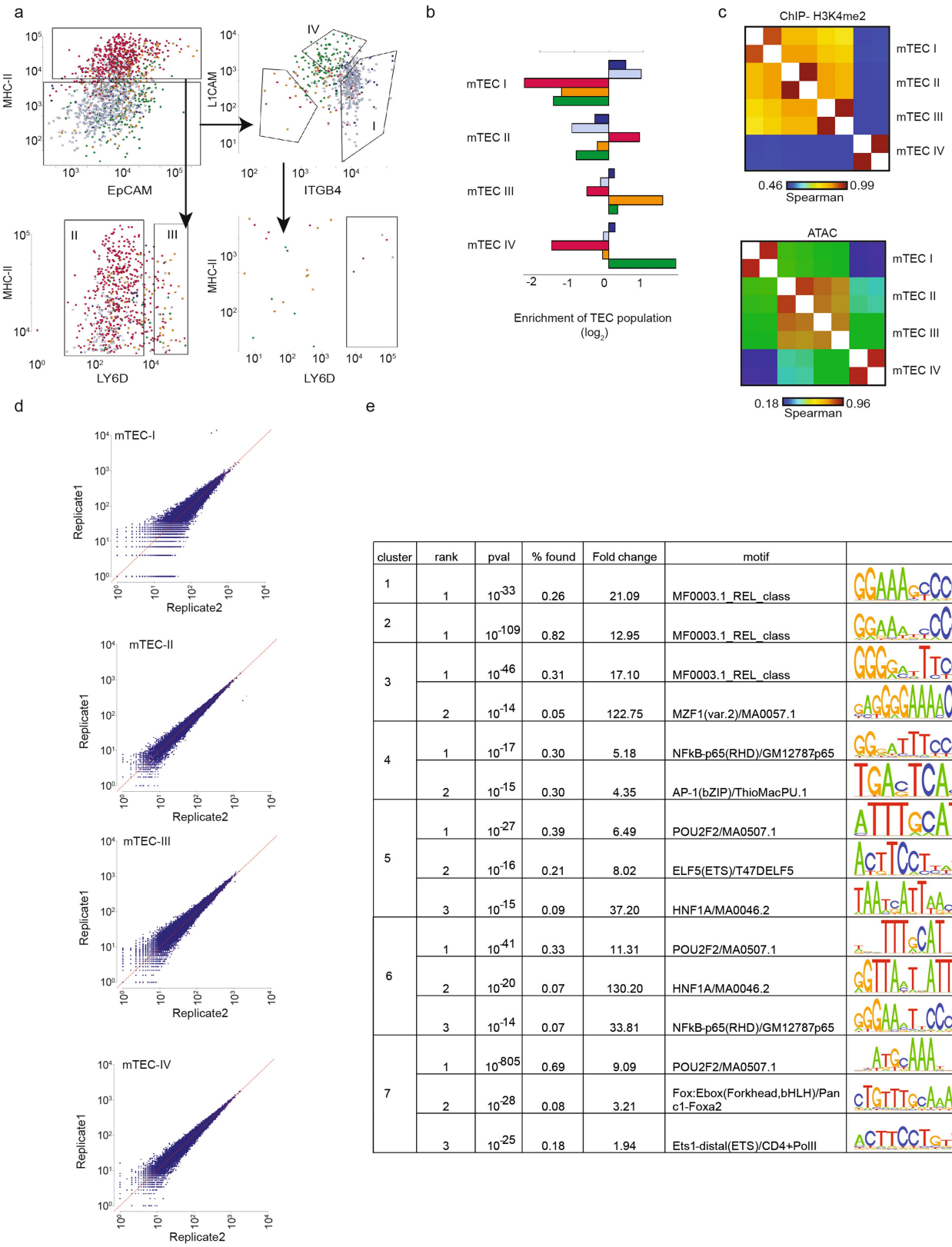


Extended Data Fig. 4 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 4 | TEC dynamics during thymus development. **a**, Flow cytometry scheme of thymic epithelial cells from different development time points. Numbers indicate fraction of $CD45^{-}EpCAM^{+}$ cells. **b**, Summary of the K ($K = 50$) nearest neighbours of embryonic cells, grouped into metacells. Neighbours of adult origin are coloured by TEC subsets, and of embryonic origin are coloured by developmental time point. Metacells with more than 20% adult neighbours were assigned to a TEC subset. **c**, Normalized mean expression of differential genes across TEC mature populations (4 weeks old) and unassigned cells from developmental time points (grey). $n = 4$ (E14.5), 3 (E18.5) and 2 (6 days) independent animals. Data are median (bars) and individual animals (dots). **d**, Differential gene expression between early cTEC (e-cTEC) from three developmental time points and the mature cTEC. Axes represent UMI count per 1,000 UMI, normalized to cell numbers. **e**, Distribution of cell cycle gene expression across cells from developing TEC is bimodal. The red line indicates the empirical proliferation threshold. **f**, Frequency of proliferating cells in the epithelial population at each developmental time point. Colour code as in **b**. **g**, Gene pairwise Spearman correlation over 2,319 e-cTEC single cells reveals three gene modules jointly expressed across embryonic early TEC and mature cTEC populations. **h**, GO annotations enrichment analysis of the three cTEC gene modules. For cell cycle, $n = 114$; antigen presentation, $n = 33$; and antigen processing, $n = 56$.

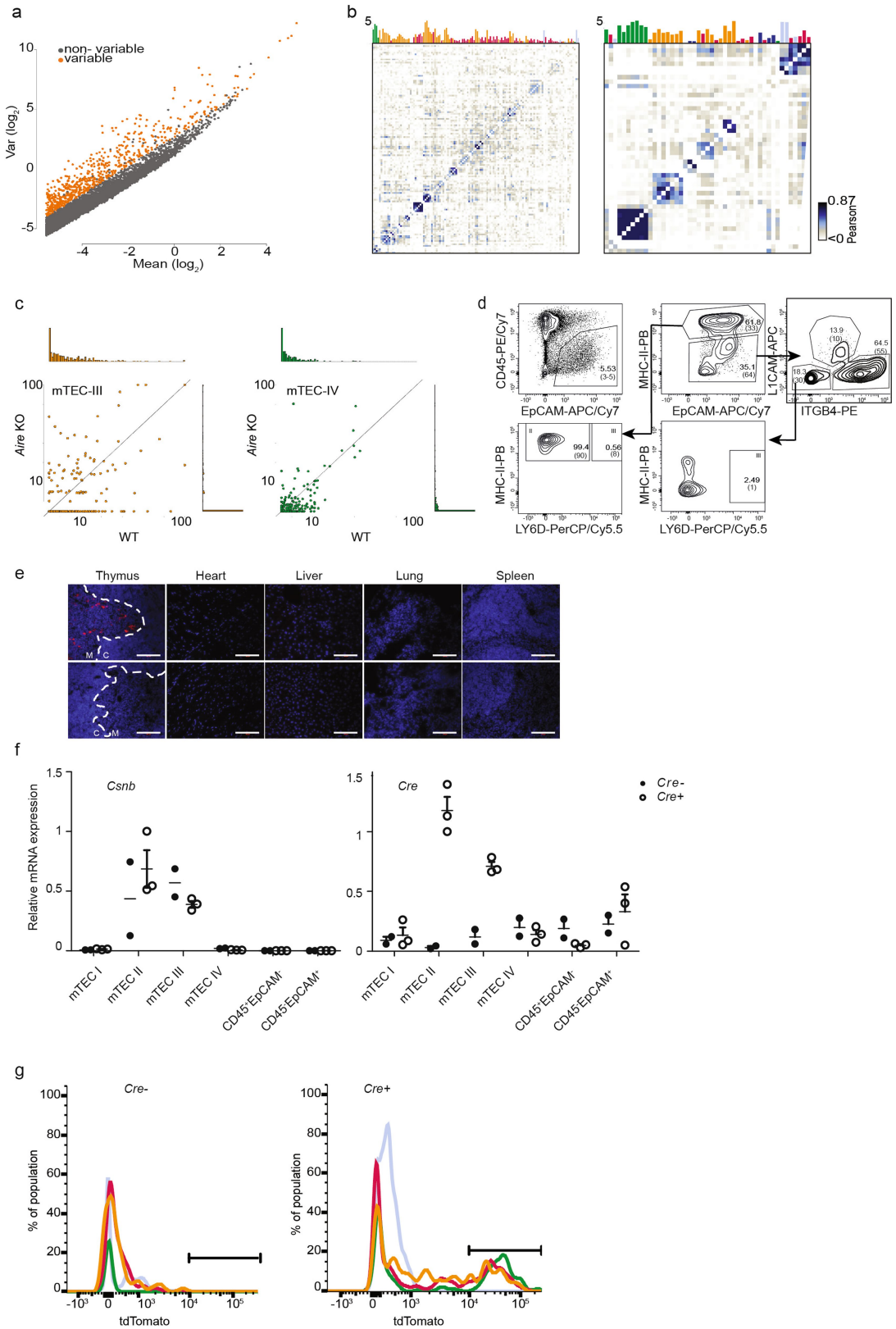
© 2018 Springer Nature Limited. All rights reserved.



cluster	rank	pval	% found	Fold change	motif	
1	1	10^{-33}	0.26	21.09	MF0003.1_REL_class	
2	1	10^{-109}	0.82	12.95	MF0003.1_REL_class	
3	1	10^{-46}	0.31	17.10	MF0003.1_REL_class	
	2	10^{-14}	0.05	122.75	MZF1(var.2)/MA0057.1	
4	1	10^{-17}	0.30	5.18	NFkB-p65(RHD)/GM12787p65	
	2	10^{-15}	0.30	4.35	AP-1(bZIP)/ThioMacPU.1	
5	1	10^{-27}	0.39	6.49	POU2F2/MA0507.1	
	2	10^{-16}	0.21	8.02	ELF5(ETS)/T47DELF5	
	3	10^{-15}	0.09	37.20	HNF1A/MA0046.2	
6	1	10^{-41}	0.33	11.31	POU2F2/MA0507.1	
	2	10^{-20}	0.07	130.20	HNF1A/MA0046.2	
	3	10^{-14}	0.07	33.81	NFkB-p65(RHD)/GM12787p65	
7	1	10^{-805}	0.69	9.09	POU2F2/MA0507.1	
	2	10^{-28}	0.08	3.21	Fox:Ebox(Forkhead,bHLH)/Panc1-Foxa2	
	3	10^{-25}	0.18	1.94	Ets1-distal(ETS)/CD4+PolII	

Extended Data Fig. 5 | Genetic and epigenetic characterization of TEC subsets. a, In silico gating of mTEC I–IV populations by index sorting measurements of surface markers. The same gating schemes were used to purify these populations by FACS (Fig. 2a). Cells are colour-coded as shown in Fig. 1. Blue, cTEC; light blue, mTEC I; red, mTEC II; yellow, mTEC III; green, mTEC IV. **b**, Relative enrichment (log_2 fold change compared to total $CD45^-EpCAM^+$ epithelial cells) of the individual mTEC I–IV subsets gated according to **a**. **c**, Heat map showing pairwise Spearman correlation of 29,472 H3K4me2 peaks (top) or ATAC-seq peaks (bottom) from mTEC I–IV sorted populations. Biological replicates for each population are shown. **d**, Scatter plots of mTEC I–IV H3K4me2 ChIP–seq peaks in biological duplicates. **e**, Summary of the most significant motifs enriched in each cluster of mTEC I–IV H3K4me2 differential peaks (Fig. 2e). *P* values are derived from binomial tests after FDR correction for multiple hypotheses. $n = 2{,}302$ differential peaks.

© 2018 Springer Nature Limited. All rights reserved.

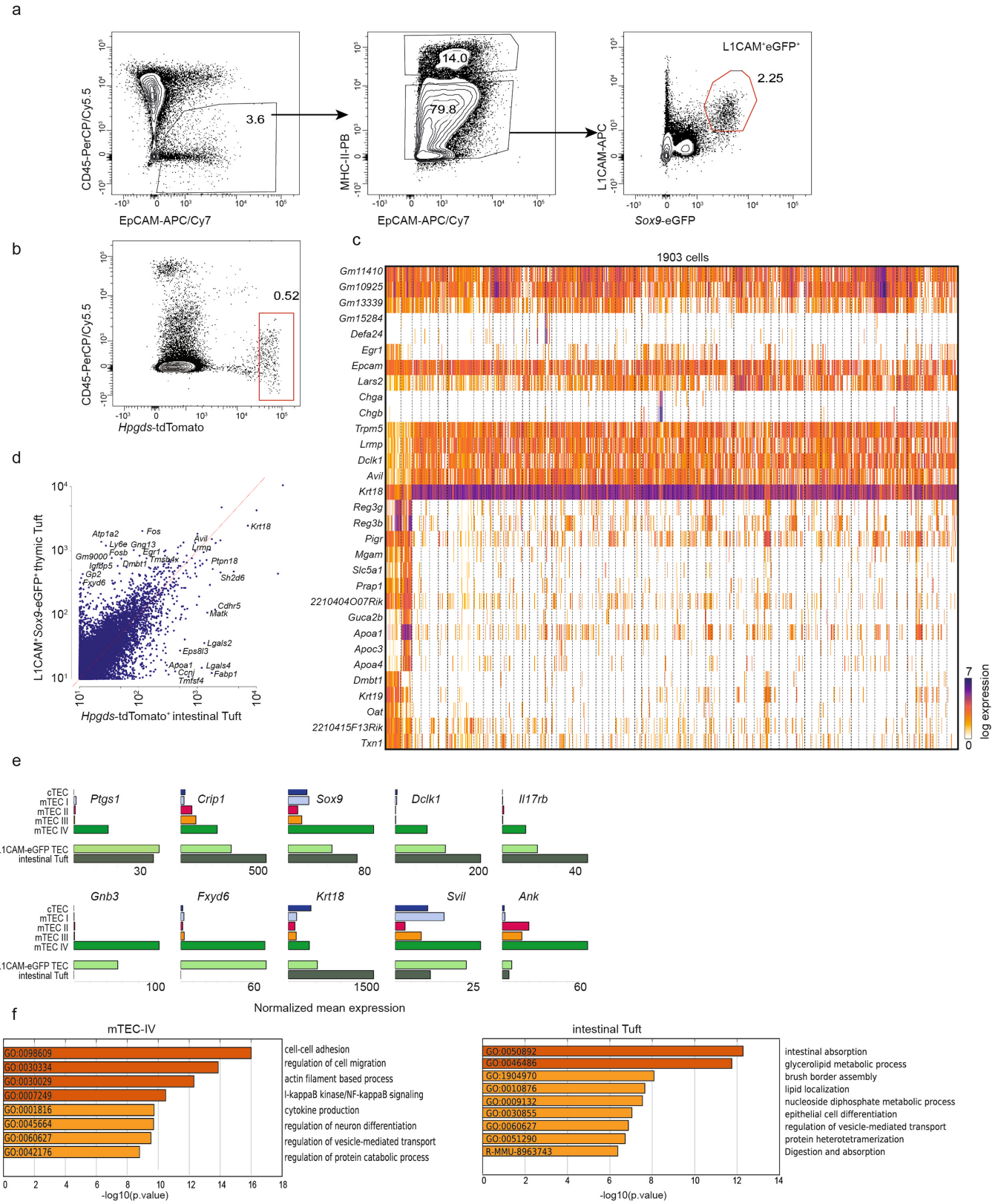


Extended Data Fig. 6 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

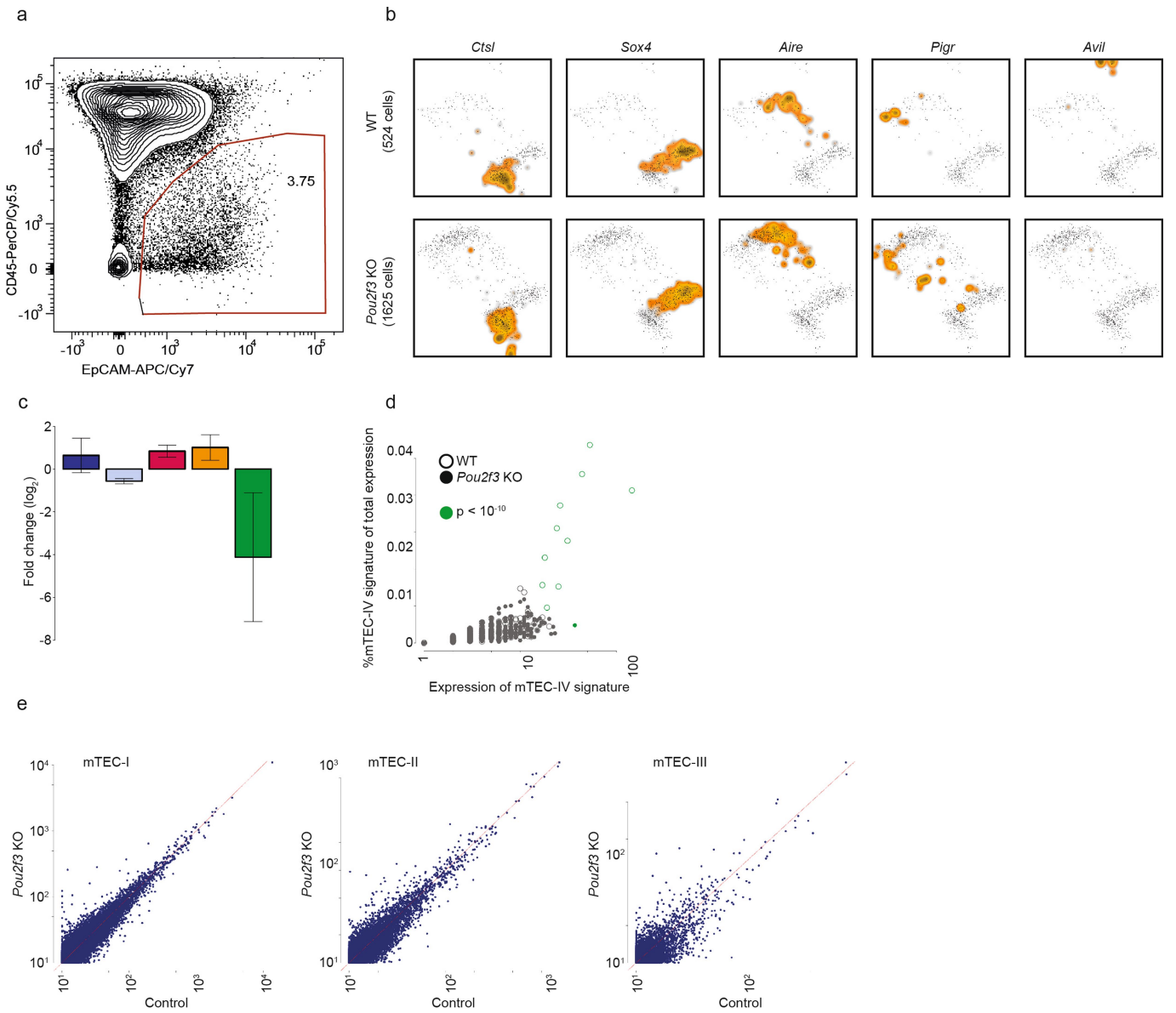
Extended Data Fig. 6 | Characterization of AIRE-dependent mTEC subsets. a, Variance of genes plotted against their mean value (genes with >50 total UMI are shown). Orange dots indicate variable genes. **b**, Pairwise Pearson gene correlations in AIRE-dependent (left) and AIRE-independent (right) TRA gene lists across 2,341 TEC single cells. Levels of differential expression (highest change of expression in cluster compared to median across all clusters) are indicated as bars; bar colours indicate cluster association to TEC population. **c**, Comparison of stochastic gene expression between *Aire* knockout and wild-type cells in mTEC III and IV populations. Marginal distribution is shown as histogram. Axes represent UMI count per 1,000 UMI, normalized to cell numbers. **d**, Flow cytometry scheme of thymic *Aire* knockout cells showing the percentage of each TEC population compared to wild-type percentage (shown in brackets). **e**, Representative immunofluorescence images of two independent experiments, for tdTomato across different organs. In the thymus, the medulla (M) and cortex (C) are separated by dashed lines, distinguished by nuclei density. Blue, DAPI. Scale bars, 100 μm. **f**, qPCR analysis of *Csnb* (left) and *cre* (right) genes in $Csnb^{cre+}Rosa26^{tdTomato}$ (cre^{+}) and $Csnb^{cre-}Rosa26^{tdTomato}$ (cre^{-}) across thymic populations. Dot plots display mean and error bars indicate s.e.m. $n = 2$ (wild type) or $n = 3$ ($Csnb^{cre+}$) biologically independent animals. **g**, Flow cytometric analysis of tdTomato expression in mTEC subsets (colours as in Fig. 1) isolated from thymi of $Csnb^{cre+}Rosa26^{tdTomato}$- or $Csnb^{cre-}Rosa26^{tdTomato}$-reporter mice.

© 2018 Springer Nature Limited. All rights reserved.



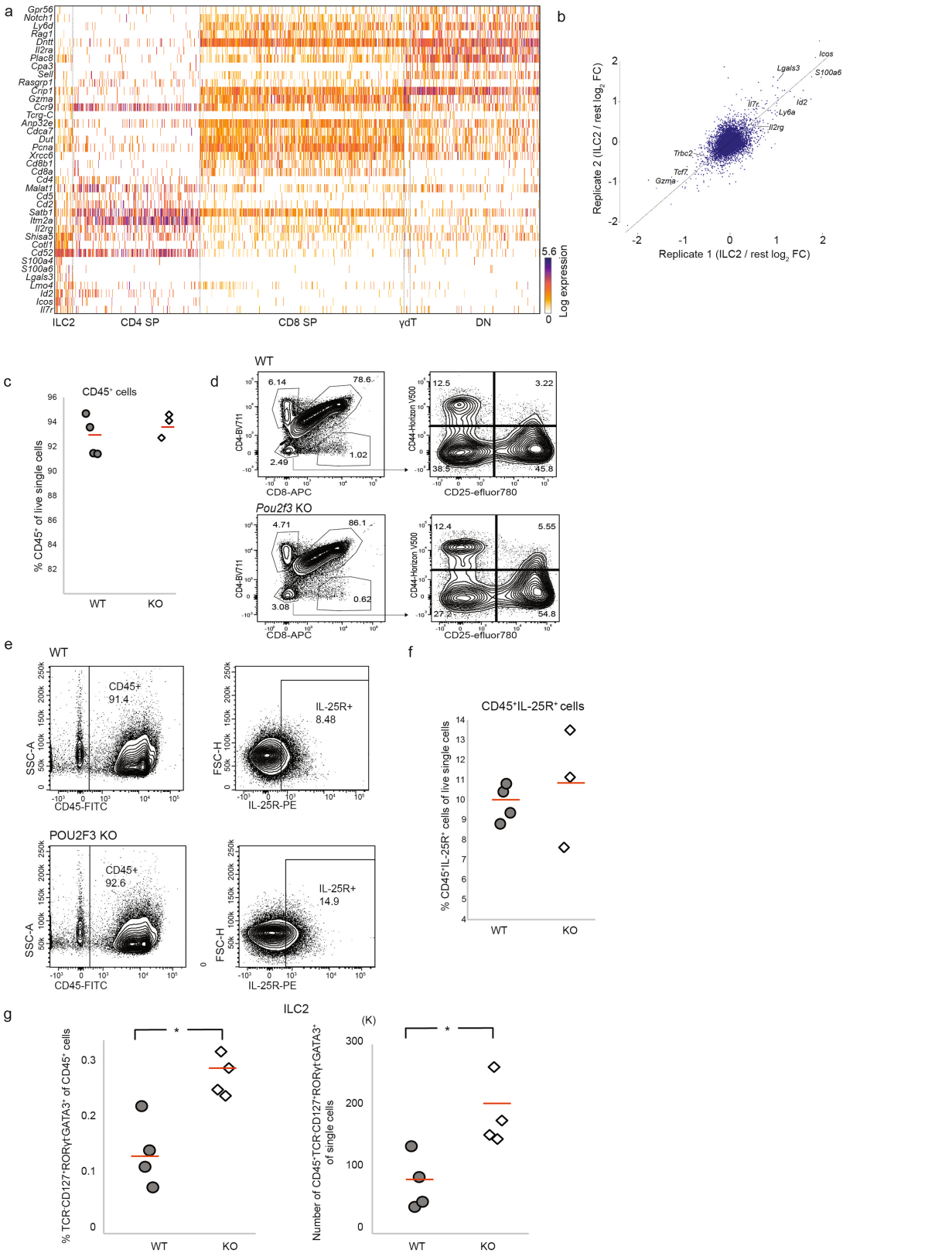
Extended Data Fig. 7 | Comparing intestinal tuft cells and the mTEC IV population. a, Flow cytometry scheme of mTEC IV cells (*Sox9*-eGFP^{+}L1CAM^{+}) sorting. **b**, Flow cytometry scheme of small intestine *Hpgds*-tdTomato^{+} tuft cell sorting. **c**, Heat map showing gene expression profiles across 1,903 intestinal tuft (*Hpgds*-tdTomato^{+}) single cells, grouped into 68 metacells. **d**, Comparison of gene expression between tuft cells isolated from small intestine (*Hpgds*-tdTomato^{+}; *x* axis) and mTEC IV cells isolated from thymus (CD45^{-}EpCAM^{+}*Sox9*-eGFP^{+}L1CAM^{+}; *y* axis). Axes represent UMI count per 1,000 UMI, normalized to cell numbers. **e**, Normalized mean expression of differential genes across TEC populations, sorted mTEC IV cells (L1CAM^{+} *Sox9*-eGFP^{+}) and intestinal Tuft (*Hpgds*-tdTomato^{+}) cells. **f**, GO annotations enrichment in differential genes (fold change >2) between intestinal tufts ($n = 634$) and mTEC IV cells (L1CAM^{+}*Sox9*-eGFP^{+}) ($n = 1{,}308$).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 8 | The transcription factor *Pou2f3* is a master regulator of mTEC IV. a, Flow cytometry scheme for sorting of EpCAM$^+$ cells from *Pou2f3* knockout thymi. **b**, Projection of representative TEC subtype-specific markers onto the two-dimensional mapping of *Pou2f3* wild-type and knockout cells to the epithelial model of Fig. 1 (see Methods). **c**, Bar plot showing $\log_2$ fold change between TEC subpopulation abundances in *Pou2f3* knockout ($n = 451$ single cells) and wild-type ($n = 1121$) mice. Error bars represent 95% confidence intervals. **d**, Pooled expression of mTEC IV genes across cells from *Pou2f3* knockout and wild-type (WT) mice. The *x* axis represents UMI count per 1,000 UMI; *y* axis represents fraction of expression from total UMI count of the cells. Green cells indicate cells classified as mTEC IV (two-sided binomial test; FDR-adjusted $P < 10^{-10}$). $n = 1{,}572$ single cells. **e**, Differential gene expression between mTEC I–III cells isolated from control (wild-type) and *Pou2f3* knockout mice. Axes represent UMI count per 1,000 UMI, normalized to cell numbers.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | mTEC IV shape the thymus immune niche. **a**, Heat map showing a metacell analysis of 3,500 CD45$^+$IL-25R$^+$ cells across five clusters. **b**, Differential gene expression in ILC2 cells from two biological replicates compared to other CD45$^+$IL-25R$^+$ cells (rest) from each replicate. Axes represent $\log_2$ fold change. **c**, Percentages of CD45$^+$ cells in *Pou2f3* knockout and wild-type thymi, determined by flow cytometry. Circles and diamonds indicate independent mice, centre line indicates the mean value. **d**, Flow cytometry analysis of cells expressing CD4, CD8, CD25 and CD44 in *Pou2f3* knockout and wild-type thymi. The experiment was repeated independently four times with similar results to confirm reproducibility. **e**, Flow cytometry sorting scheme of thymic CD45$^+$IL-25R$^+$ cells from *Pou2f3* knockout and wild-type thymi. **f**, Percentage of CD45$^+$IL-25R$^+$ cells in *Pou2f3* knockout and wild-type thymi. Circles and diamonds indicate independent mice, centre line indicates the mean value. **g**, Percentages (left) and numbers (right) of ILC2 (Lin$^-$TCR$^-$CD127$^+$GATA3$^+$Rorγt$^-$) cells within the single-cell gate in *Pou2f3* knockout and wild-type mice. Circles and diamonds indicate independent mice, centre line indicates the mean value. A one-tailed Student's *t*-test was used for the comparison, $^*P < 0.05$.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Regions of enhancer enrichment peaks

a.

Sample	Replicate	number of iChiP aligned reads	number of IDR peaks
mTEC-I	a	13831980	7037
mTEC-I	b	3359743	
mTEC-II	a	32464771	23950
mTEC-II	b	29099226	
mTEC-III	a	17032161	19322
mTEC-III	b	17710805	
mTEC-IV	a	18818509	22758
mTEC-IV	b	17910967	

b.

Sample	Replicate	number of ATAC aligned reads	Number of peaks
mTEC-I	a	12521131	44288
mTEC-I	b	10665995	39040
mTEC-II	a	6626821	19008
mTEC-II	b	8633041	27739
mTEC-III	a	10899189	10525
mTEC-III	b	2387938	19507
mTEC-IV	a	8388701	30017
mTEC-IV	b	13972276	38108

a, H3K4me2 iChIP–seq peak calling. **b**, ATAC-seq peak calling.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0345-2

Thymic tuft cells promote an IL-4-enriched medulla and shape thymocyte development

Corey N. Miller[1,2], Irina Proekt[1,2], Jakob von Moltke[2,3,4,9], Kristen L. Wells[5], Aparna R. Rajpurkar[5], Haiguang Wang[6], Kristin Rattay[7,10], Imran S. Khan[1,2,11], Todd C. Metzger[1,2,12], Joshua L. Pollack[2,13], Adam C. Fries[8], Wint W. Lwin[1,2], Eric J. Wigton[3], Audrey V. Parent[1,2], Bruno Kyewski[7], David J. Erle[2], Kristin A. Hogquist[6], Lars M. Steinmetz[5], Richard M. Locksley[2,3,4] & Mark S. Anderson[1,2,3]*

The thymus is responsible for generating a diverse yet self-tolerant pool of T cells[1]. Although the thymic medulla consists mostly of developing and mature AIRE$^+$ epithelial cells, recent evidence has suggested that there is far greater heterogeneity among medullary thymic epithelial cells than was previously thought[2]. Here we describe in detail an epithelial subset that is remarkably similar to peripheral tuft cells that are found at mucosal barriers[3]. Similar to the periphery, thymic tuft cells express the canonical taste transduction pathway and IL-25. However, they are unique in their spatial association with cornified aggregates, ability to present antigens and expression of a broad diversity of taste receptors. Some thymic tuft cells pass through an *Aire*-expressing stage and depend on a known AIRE-binding partner, HIPK2, for their development. Notably, the taste chemosensory protein TRPM5 is required for their thymic function through which they support the development and polarization of thymic invariant natural killer T cells and act to establish a medullary microenvironment that is enriched in the type 2 cytokine, IL-4. These findings indicate that there is a compartmentalized medullary environment in which differentiation of a minor and highly specialized epithelial subset has a non-redundant role in shaping thymic function.

Medullary thymic epithelial cells (mTECs) develop through a coordinated pathway in which AIRE-negative MHC-II^{low} precursors receive inductive signals, including RANKL and CD40L, mature into MHC-II^{high}CD80^{high} (mTEC^{high}) AIRE- and tissue-specific antigen (TSA)-expressing cells, and promote central tolerance through negative selection and generation of thymically derived regulatory T cells[2]. We developed a lineage-tracing model in which *Aire*-expressing cells were inducibly labelled with a fluorescent reporter (*Aire*^{creERT2};*Rosa26*$^{CAG\text{-}stopflox\text{-}tdTomato}$ or *Aire*^{creERT2};*Rosa26*$^{CAG\text{-}stopflox\text{-}zsGreen}$; inducible AIRE lineage trace, iALT)[4]. In iALT mice that were treated with a single dose of tamoxifen, four mTEC subsets could be distinguished based on MHC-II expression and RFP, including a post-*Aire* (MHC-II^{low}RFP$^+$) population[4].

To investigate mTEC heterogeneity, we used fluorescence-activated cell sorting (FACS) to sort iALT-labelled mTECs and examined their transcriptomes (Fig. 1a, b). Bulk RNA sequencing (RNA-seq) revealed the highest levels of *Aire* and TSA transcripts in the early-*Aire* (MHC-II^{high}RFP^{low}) and late-*Aire* (MHC-II^{high}RFP^{high}) subsets and unsupervised hierarchical clustering showed that these populations were the most related to one another (Extended Data Fig. 1a, b). In the post-*Aire* subset, two distinct transcriptional signatures emerged. The first was enriched for markers of the soft cornified epithelial pathway[5], including the late-stage cytokeratin, *Krt10* (Fig. 1b, c). This is consistent with the observation of cornified bodies within the human thymus, known as Hassall's corpuscles, and recent reports of cornified markers within the mouse thymus[6–10]. Immunofluorescence analysis confirmed robust expression of KRT10 protein in the thymus of wild-type mice and confocal microscopy images showed the distinctive morphology of the medullary KRT10$^+$ structures (KRT10 bodies) (Extended Data Fig. 2a).

The second transcriptional signature included genes that are associated with an epithelial subset called tuft cells[11] (Fig. 1b, d). Recent reports have shown that these cells have a non-redundant chemosensory role in the intestine in which they orchestrate a feedforward loop driving the type 2 response to helminths and protozoa[12–14]. Tuft cells are notable for their expression of the canonical taste transduction pathway (that is, *Gnat3*, *Plcg2* or *Plcb2* and *Trpm5*), choline O-acetyltransferase (*Chat*), doublecortin-like kinase 1 (*Dclk1*) and as the sole source of the barrier-associated cytokine, *Il25*[3,15] (Fig. 1d). The downstream cation channel, *Trpm5*, is required for tuft function in the intestine, but the upstream sensory receptor(s) remains unknown, although some peripheral tuft cells express a limited repertoire of type-II taste receptors from the bitter ligand family (TAS2R)[16,17]. Flow cytometric analysis of mTECs demonstrated that approximately 10% of mTECs in the thymus of adult C57BL/6 mice were DCLK1^{bright} and immunofluorescence staining showed that DCLK1^{bright} mTECs were distributed throughout the medulla (Fig. 1e and Extended Data Fig. 2a, b). These cells frequently had a bulbous morphology, narrow protruding base, and were often grouped into small multicellular clusters[15] (Fig. 1f and Extended Data Fig. 2b). Unexpectedly, DCLK1^{bright} cells were closely associated with KRT10 bodies and quantitative image analysis confirmed that they were significantly more likely to be adjoining KRT10 surfaces than predicted by random chance (Fig. 1g and Extended Data Fig. 3a–d). In the human thymus, medullary DCLK1^{bright} cells regularly abutted Hassall's corpuscles, and were 3.5% of CD45$^-$EPCAM$^+$ TECs (Extended Data Fig. 4a, b).

Although the presence of CHAT and GNAT3, and the expression of several TAS2R family members have previously been reported in AIRE$^-$ mTECs by Panneck et al.[18] and Soultanova et al.[19], the importance of these observations has been unclear. Therefore, we wanted to understand how closely related DCLK1^{bright} mTECs were to intestinal tuft cells. We used the Flare25 IL-25–RFP reporter to confirm that DCKL1^{bright} mTECs expressed IL-25[12]. RFP$^+$ cells were dispersed throughout the medulla and colocalized with DCLK1, confirming tonic IL-25 production, and were approximately 7–10% of mTECs (Fig. 2a and Extended Data Fig. 5a, b). We then used the Flare25 reporter to sort small-intestinal enterocytes (RFP$^-$), small-intestinal tuft cells (RFP$^+$)

[1]Diabetes Center, University of California, San Francisco, San Francisco, CA, USA. [2]Department of Medicine, University of California, San Francisco, San Francisco, CA, USA. [3]Department of Microbiology and Immunology, University of California, San Francisco, San Francisco, CA, USA. [4]Howard Hughes Medical Institute, University of California, San Francisco, San Francisco, CA, USA. [5]Department of Genetics, Stanford University School of Medicine, Stanford, CA, USA. [6]Center for Immunology, Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN, USA. [7]Division of Developmental Immunology, Tumor Immunology Program, German Cancer Research Center (DKFZ), Heidelberg, Germany. [8]Biological Imaging Development Center and Department of Pathology, University of California, San Francisco, CA, USA. [9]Present address: Department of Immunology, University of Washington, Seattle, WA, USA. [10]Present address: Division of Immunology, Harvard Medical School, Boston, MA, USA. [11]Present address: Department of Pediatrics, University of California, San Francisco, San Francisco, CA, USA. [12]Present address: Bristol-Myers Squibb, Sunnyvale, CA, USA. [13]Present address: Pionyr Immunotherapeutics, San Francisco, CA, USA. *e-mail: mark.anderson@ucsf.edu

© 2018 Springer Nature Limited. All rights reserved.

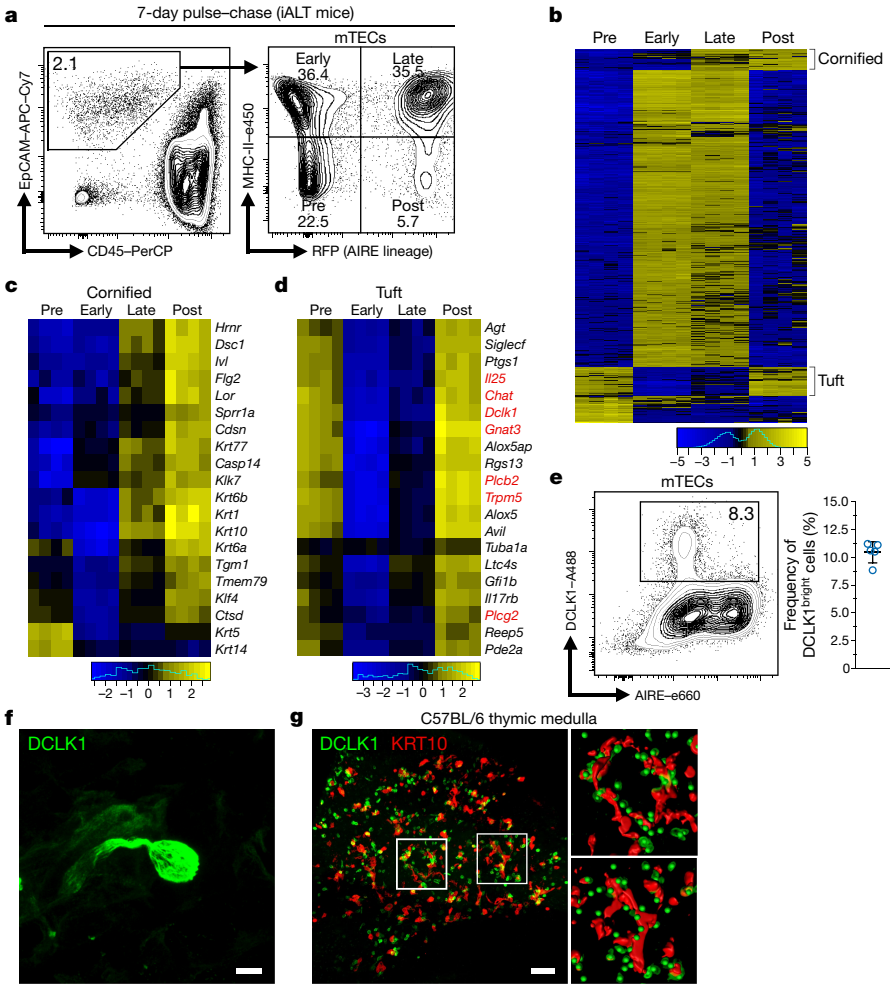


Fig. 1 | Tuft-like cells are closely associated with cornified bodies in the thymic medulla. a, Gating of mTEC subsets within $CD11c^{-}CD45^{-}EPCAM^{+}$ thymic epithelial cells. Sorted in quadruplicate for RNA-seq (12 pooled thymi per replicate, $n = 4$ sorted replicates). **b**, Heat map of differentially expressed genes (false-discovery rate (FDR) < 0.01 and $|\text{fold change}| > 8$). **c**, **d**, Selected genes from regions marked 'cornified' (**c**) or 'tuft' (**d**). $\log_2$ fold change in expression relative to mean expression levels. **e**, DCLK1 intracellular staining in mTECs. Data are mean ± s.d. $n = 5$ mice; 3 independent experiments. **f**, Confocal maximum projection of a $DCLK1^{bright}$ cell. Scale bar, 5 μm. $n = 5$ mice, 3 independent experiments. **g**, Left, confocal maximum projection ($z = 77$ μm) of a medullary region at low magnification. Right, regions of interest (white squares) with KRT10 converted to surfaces and DCLK1 converted to centre of intensity coordinates. Scale bar, 100 μm. $n = 3$ thymic slices, 2 independent experiments.

and thymic RFP^{+} cells for bulk RNA-seq and compared expression of 20 tuft markers[11,20] (Fig. 2b). Of these, 18 were highly expressed in small-intestinal tuft cells and thymic RFP^{+} mTECs, whereas none were expressed in RFP^{-} small-intestinal enterocytes. Notably, only RFP^{+} mTECs strongly expressed *Gnat3*, the G_{α} subunit of gustducin and upstream mediator of the taste signal transduction pathway, whereas both intestinal and thymic RFP^{+} populations expressed *Trpm5*[21,22] (Fig. 2b). FACS and RT–qPCR analysis confirmed that RFP^{+} mTECs were the dominant source of *Il25* and *Trpm5* mRNA (Extended Data Fig. 5c, d). Finally, $DCLK1^{bright}$ mTECs were also observed to be $KRT8/18^{+}$, consistent with peripheral tuft cells (Extended Data Fig. 5e). These data confirm that thymic $DCLK1^{bright}$ cells represent a novel subset of tuft cells.

Despite shared core transcriptional features, we hypothesized that thymic tuft cells might differ from their peripheral counterparts in important ways, such as the ability to present antigens. Small-intestinal tuft cells expressed MHC-I and β_2-microglobulin, however, only thymic tuft cells expressed MHC-II genes and CD74 (Fig. 2c and Extended Data Fig. 5f–h). Expression of MHC-II on the surface of thymic tuft cells was confirmed by flow cytometry, although overall surface levels were lower than $AIRE^{+}$ $mTEC^{high}$ cells (Figs. 1a, 2d). Whereas small-intestinal tuft cells did not express any of the known TAS2R family members, at the bulk RNA level, thymic tuft cells expressed a broad array of these receptors (Fig. 2e and Extended Data Fig. 5i). When examined at the single-cell level, *Tas2r* genes tended to be coexpressed in single cells, although *Tas2r* expression was far more heterogeneous than *Dclk1* or *Trpm5* expression (Fig. 2f and Extended Data Figs. 6a, 7b). When coexpression of *Tas2r* genes was compared to the background gene set, many *Tas2r* gene pairs were significantly highly correlated (empirical $P < 0.05$), suggesting that expression is not random and subsets of thymic tuft cells may be coordinated in their chemosensory specificities (Extended Data Fig. 6a–c). Cells that expressed TAS2R family members also tended to express *Gnat3* and coexpress 3–8 TAS2R family members (Fig. 2f and Extended Data Fig. 7b). These unique properties of thymic tuft cells raised the possibility that they participate in antigen presentation to developing thymocytes and may use their chemosensory abilities as part of their thymic function.

Notably, $Aire^{-/-}$ animals had normal numbers of thymic tuft cells[23] (Fig. 3a and Extended Data Fig. 7a). After tamoxifen treatment for 10 days, thymi from iALT × Flare25 reporter mice had an RFP^{+} single-positive population as well as a clear double-positive $RFP^{+}zsGreen^{+}$ mTEC population, confirming that some thymic tuft cells pass through an *Aire*-expressing stage and also suggesting that an *Aire*-independent pathway exists (Fig. 3b). This is consistent with the thymic tuft signature appearing in both the pre- and post-*Aire* populations (Fig. 1b). At the single-cell level, $Aire^{-/-}$ thymic tuft cells showed

© 2018 Springer Nature Limited. All rights reserved.

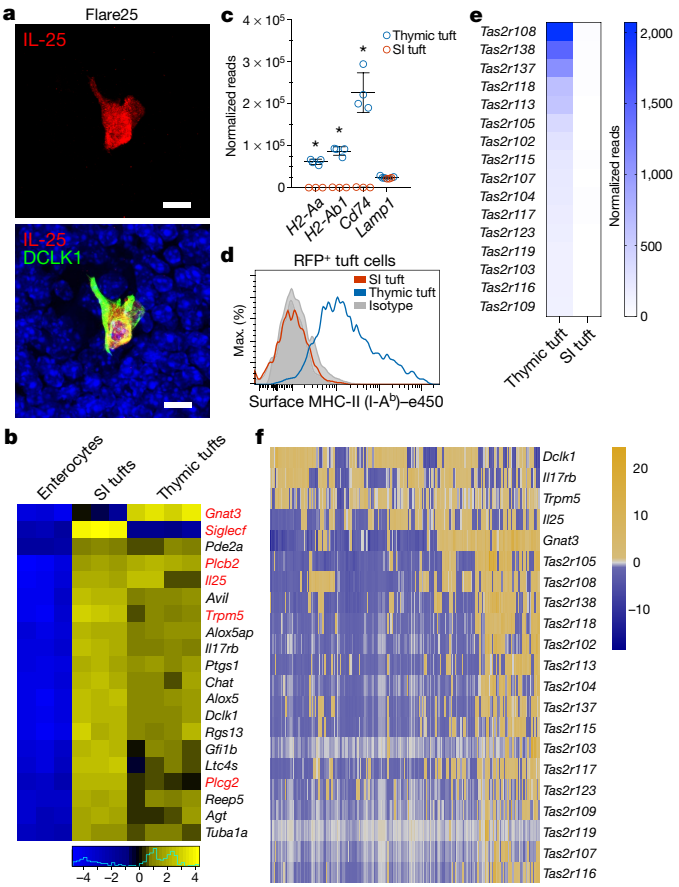


Fig. 2 | Thymic DCLK1^{bright} cells are a novel subset of tuft cells. **a**, Colocalization of DCLK1 and RFP (IL-25) in the thymus of a Flare25 mouse. Confocal maximal projection. Scale bars, 5 μm. $n = 3$ mice, 2 independent experiments. **b**, Differential expression of 20 tuft markers comparing sorted enterocytes and small-intestinal tuft cells ($n = 3$ mice) and thymic tuft cells ($n = 4$ mice) from Flare25 mice. $\log_2$ fold change in expression relative to mean expression levels. **c**, Expression (normalized reads from **b**) of MHC-II genes in small-intestinal and thymic tuft cells. Data are mean ± s.d. *FDR < 0.00001. *Lamp*, FDR = 1. **d**, MHC-II (I-A^b) surface staining on small-intestinal and thymic tuft cells. $n = 3$ mice, 2 independent experiments. **e**, Heat map of select *Tas2r* genes comparing expression (normalized reads from **b**) between thymic and small-intestinal tuft cells. **f**, Single-cell RNA-seq of RFP$^+$ (IL-25$^+$) thymic tuft cells. Heat map shows gene expression of selected tuft and *Tas2r* genes. Columns are single cells ($n = 195$), values are mean-centred log-normalized.

a similar pattern of expression of core markers and TAS2R family members as in wild-type cells (Fig. 3c and Extended Data Figs. 7b–e, 8a, b). However, principal component analysis showed that the majority of *Aire*$^{-/-}$ thymic tuft cells were mostly separated from wild-type cells (Fig. 3d). Notably, *Gnat3* was a driver of the variation observed in principal component 1 and showed heterogeneous expression across both wild-type and *Aire*$^{-/-}$ cells (Fig. 3d). This separation was supported by quantitative image analysis of *Aire*$^{-/-}$ thymic sections that had a marked increase in the frequency and size of tuft cell aggregates (Fig. 3e). Notably, there was a marked decrease in KRT10 bodies in the thymus of *Aire*$^{-/-}$ mice; this may contribute to the observed disorganization of the thymic tuft compartment in these mice[23] (Fig. 3e, f and Extended Data Figs. 3c, 7f).

HIPK family members mediate signal integration and cell-fate decisions during the differentiation of complex tissues. *Hipk2* has been shown to be an AIRE-binding partner, however, mTEC-specific ablation of *Hipk2* had only a small effect on mTEC^{high} gene expression, including TSAs[24]. Instead, there was decreased expression of a subset of genes within MHC-II^{low} mTECs, including the majority of tuft-associated genes. As such, we hypothesized that HIPK2 might promote differentiation of thymic tuft cells. Indeed, *Foxn1*^{cre}*Hipk2*$^{fl/fl}$ mice had a profound reduction in the frequency of MHC-II^{low}DCLK1^{bright} cells with a concomitant increase in MHC-II^{low} and MHC-II^{high}DCLK1^{dim} cells, suggesting a developmental blockade (Fig. 3g, h and Extended Data Fig. 7g). Therefore, we describe a novel *Hipk2*-dependent mechanism for tuft cell differentiation.

Because tuft cells drive intestinal type 2 immune responses, we hypothesized that they might promote a type 2 microenvironment in the thymus. Previous work has shown that type 2 invariant natural killer T (NKT2) cells are the dominant source of interleukin-4 (IL-4) in the medulla[25,26]. We used the BALB/cByJ × *Aire*^{DTR} C57BL/6 F_1 model, characterized by increased numbers of thymic NKT2 cells and IL-4-dependent EOMES$^+$CD8$^+$ single-positive thymocytes, and ablated post-*Aire* tuft cells with a nine-day course of diphtheria toxin[26] (Extended Data Fig. 9a, b). This resulted in a selective reduction in the number of NKT2 cells and EOMES$^+$CD8$^+$ single-positive thymocytes[26] (Extended Data Fig. 9c, d). Importantly, there was a marked decrease in IL-4 production by the remaining NKT2 cells (Fig. 4a). Because this model ablates both mTEC^{high} and post-*Aire* mTECs, we exploited the requirement for the transcription factor *Pou2f3* for peripheral tuft cell development[27]. As in the periphery, we found no thymic tuft cells in *Pou2f3*$^{-/-}$ mice but a normal frequency of AIRE$^+$ mTECs, grossly normal thymic architecture and normal numbers of double-negative (CD4$^-$CD8$^-$), double-positive (CD4$^+$CD8$^+$) and single-positive (CD4$^+$or CD8$^+$) thymocytes (Fig. 4b and Extended Data Fig. 10a–c). However, there were fewer TCRβ^{int}CD1d$^+$ invariant natural killer T (iNKT) cells and, as in the F_1 model, there was a marked decrease in NKT2 cells and EOMES$^+$CD8$^+$ single-positive thymocytes (Fig. 4c, d and Extended Data Fig. 10d–f). Notably, *Trpm5*$^{-/-}$ mice phenocopied the thymocyte perturbations of *Pou2f3*$^{-/-}$ mice, demonstrating that *Trpm5* is required for thymic tuft cell function (Fig. 4c, d). Finally, the number of splenic NKT2 cells was selectively reduced in *Pou2f3*$^{-/-}$ mice, mirroring the thymic defect (Fig. 4e and Extended Data Fig. 10g). Together, these data show that thymic tuft cells are critical for the development and intrathymic function of NKT2s, require components of the canonical taste chemosensory pathway and contribute to the availability of thymic IL-4.

Given that thymic tuft cells express surface MHC-II, they could present self-antigens not covered by the mTEC^{high} population. To directly test this, we used a thymic transplant model in which thymi from C57BL/6 or *Pou2f3*$^{-/-}$ neonatal mice were transplanted into athymic *Foxn1*$^{-/-}$ nude mice, allowed to reconstitute, and immunized with IL-25 protein. Remarkably, only *Pou2f3*$^{-/-}$ transplant recipients, which had peripheral tuft cells but lacked antigen expression by thymic tuft cells, generated an IL-25-specific antibody response (Fig. 4f). Therefore, thymic tuft cells are required to enforce tolerance to a canonical tuft-restricted antigen and probably serve this purpose for other such proteins.

Our findings reveal a previously undescribed microanatomical niche consisting of terminally differentiated cornified epithelial cell aggregates and thymic tuft cells. Thymic tuft cells are a novel tuft subset that have characteristics of both mTECs and peripheral tuft cells. Their development is dependent on the transcription factors *Pou2f3* and *Hipk2*, a known AIRE-binding partner. We demonstrate a role for thymic tuft cells in the establishment of an IL-4-enriched medullary microenvironment through NKT2-mediated IL-4 production. Thymic tuft cells depend on *Trpm5* for their function, strongly indicating that they respond to local environmental cues through their chemosensory abilities. In summary, our findings provide a novel source of antigen that is required for the establishment of tolerance and an upstream mediator of thymic IL-4 production through an innate-like (iNKT) lymphoid pathway. The former extends our knowledge of an established role for the thymus in immune function and the latter underscores an emerging role for adaptive effector cytokines beyond peripheral immune responses, including thymocyte development.

© 2018 Springer Nature Limited. All rights reserved.

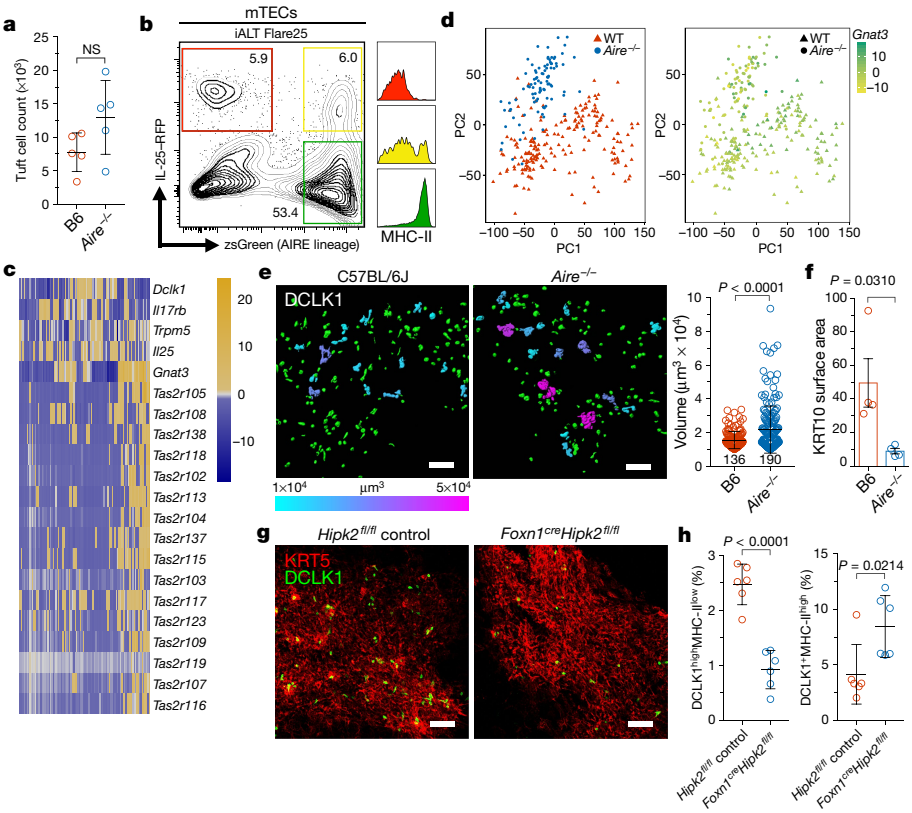


Fig. 3 | *Aire* is expressed by some thymic tuft cells but is not required for their development. a, DCLK1^{bright} mTEC counts. $n = 5$ mice per genotype, 3 independent experiments. **b**, Left, mTECs from iALT;Flare25 double-reporter mice after 10-day tamoxifen treatment. Right, MHC-II (I-A^b) surface expression within each gated population. $n = 5$ mice, 2 independent experiments. **c**, Single-cell RNA seq of RFP$^+$ (IL-25$^+$) *Aire*$^{-/-}$ thymic tuft cells. Heat map shows gene expression of selected tuft and *Tas2r* genes. Columns are single cells ($n = 95$), values are mean-centred log-normalized. **d**, Left, principal component analysis of single-cell RNA-seq data from RFP$^+$ tuft cells. Right, data from the left plot with *Gnat3* expression overlaid. **e**, Left, DCLK1 staining with signal converted to Imaris surfaces; volumes >5× mean single-cell volume (1×10^4 µm^3) are pseudocoloured according to size. Right, pooled data. $n = 3$ mice per genotype. **f**, KRT10 surface area in thymic medulla (per unit area). Three regions were analysed per thymus. $n = 4$ mice per genotype, 2 independent experiments. **g**, **h**, Analysis of DCLK1^{bright} cells in *Hipk2*$^{fl/fl}$ controls or *Foxn1*^{cre}*Hipk2*$^{fl/fl}$ thymus. **g**, Confocal imaging of immunofluorescence staining. **h**, Frequencies of DCLK1^{high}MHC-II^{low} or DCLK1$^+$MHC-II^{high} mTECs from flow analysis. $n = 6$ mice, 3 independent experiments. **e**, **g**, Scale bars, 100 µm. **a**, **e**, **f**, **h**, Data are mean ± s.d.; unpaired, parametric, two-tailed Student's *t*-test.

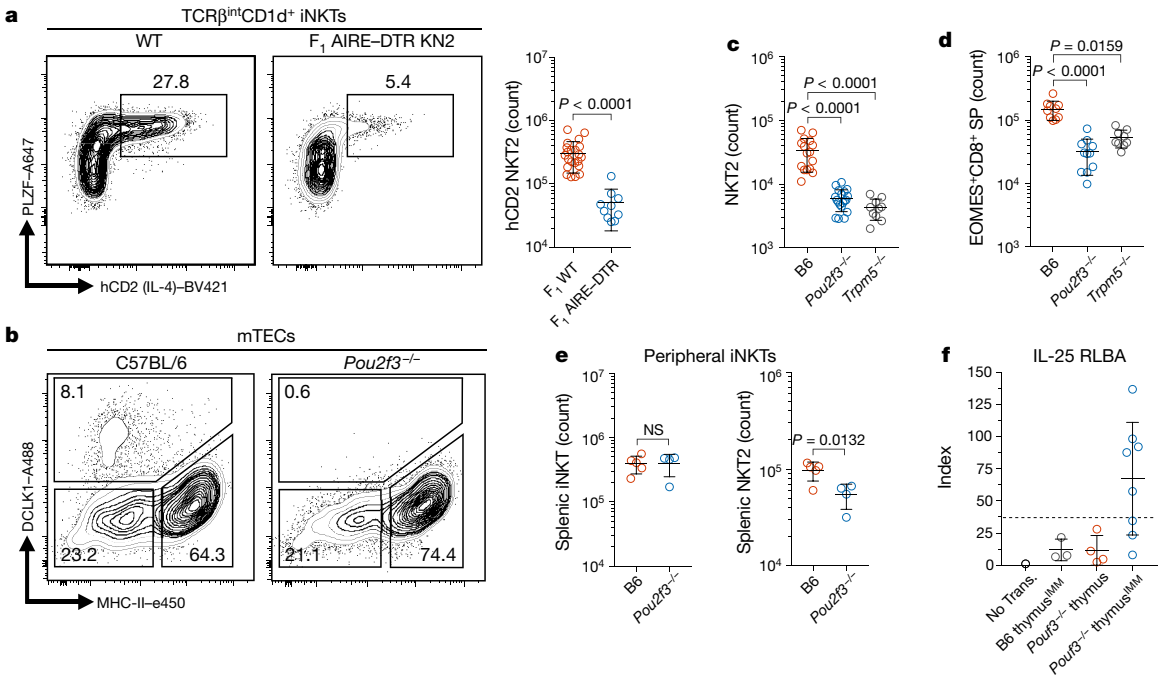


Fig. 4 | Thymic tuft cells promote an IL-4-enriched medullary microenvironment and enforce tolerance. a, Thymus cells from diphtheria toxin-treated *Kn2* BALB/cByJ × *Aire*^{DTR} C57BL/6 F$_1$ mice ($n = 10$ mice) or non-transgenic controls ($n = 22$ mice) gated on TCRβ^{int} CD1d$^+$ iNKT cells. Right, counts of IL-4-producing (human (h)CD2$^+$) NKT2s (PLZF$^+$). WT, wild type. **b**, DCLK1 intracellular staining in mTECs. $n = 5$ mice; 3 independent experiments. **c**, Thymic NKT2 (TCRβ^{int}CD1d$^+$PLZF$^+$RORγt$^-$) cell counts in C57BL/6 (B6) ($n = 15$ mice), *Pou2f3*$^{-/-}$ ($n = 19$ mice) and *Trpm5*$^{-/-}$ ($n = 10$ mice) mice. **d**, Thymic EOMES$^+$TCRβ$^+$CD8$^+$ single-positive (SP) cell counts in C57BL/6 ($n = 10$ mice), *Pou2f3*$^{-/-}$ ($n = 11$ mice) and *Trpm5*$^{-/-}$ ($n = 9$ mice) mice. **e**, Splenic total iNKT and NKT2 cell counts in C57BL/6 ($n = 5$ mice) and *Pou2f3*$^{-/-}$ ($n = 4$ mice) mice. **f**, Serum anti-IL-25 autoantibody indices in nude mice transplanted with C57BL/6 ($n = 3$ mice) or *Pou2f3*$^{-/-}$ ($n = 12$ mice) neonatal thymus and immunized with IL-25 protein. IMM, immunized. Dashed line, the average of C57BL/6 autoantibody index values plus three standard deviations. RLBA, radioligand binding assay. **a**, **c**–**f**, Data are mean ± s.d. **a**, **e**, Unpaired, parametric, two-tailed Student's *t*-test. **c**, **d**, One-way, non-parametic ANOVA (Kruskal–Wallis test). **a**, **c**, **d**, Pooled from three independent experiments.

© 2018 Springer Nature Limited. All rights reserved.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0345-2.

Received: 13 October 2017; Accepted: 23 May 2018;
Published online 18 July 2018.

1. Kyewski, B. & Klein, L. A central role for central tolerance. *Annu. Rev. Immunol.* **24**, 571–606 (2006).
2. Takahama, Y., Ohigashi, I., Baik, S. & Anderson, G. Generation of diversity in thymic epithelial cells. *Nat. Rev. Immunol.* **17**, 295–305 (2017).
3. Gerbe, F. & Jay, P. Intestinal tuft cells: epithelial sentinels linking luminal cues to the immune system. *Mucosal Immunol.* **9**, 1353–1359 (2016).
4. Metzger, T. C. et al. Lineage tracing and cell ablation identify a post-Aire-expressing thymic epithelial cell population. *Cell Rep.* **5**, 166–179 (2013).
5. Eckhart, L., Lippens, S., Tschachler, E. & Declercq, W. Cell death by cornification. *Biochim. Biophys. Acta* **1833**, 3471–3480 (2013).
6. Derbinski, J. et al. Promiscuous gene expression in thymic epithelial cells is regulated at multiple levels. *J. Exp. Med.* **202**, 33–45 (2005).
7. White, A. J. et al. Lymphotoxin signals from positively selected thymocytes regulate the terminal differentiation of medullary thymic epithelial cells. *J. Immunol.* **185**, 4769–4776 (2010).
8. Odaka, C. et al. TGF-β type II receptor expression in thymic epithelial cells inhibits the development of Hassall's corpuscles in mice. *Int. Immunol.* **25**, 633–642 (2013).
9. Hassall, A. H. *The Microscopic Anatomy of the Human Body, in Health and Disease* Vol. 2, Plate LXI, Figure 10 (Samuel Highley, London, 1849).
10. Farr, A. G., Dooley, J. L. & Erickson, M. Organization of thymic medullary epithelial heterogeneity: implications for mechanisms of epithelial differentiation. *Immunol. Rev.* **189**, 20–27 (2002).
11. Gerbe, F., Legraverend, C. & Jay, P. The intestinal epithelium tuft cells: specification and function. *Cell. Mol. Life Sci.* **69**, 2907–2917 (2012).
12. von Moltke, J., Ji, M., Liang, H.-E. & Locksley, R. M. Tuft-cell-derived IL-25 regulates an intestinal ILC2-epithelial response circuit. *Nature* **529**, 221–225 (2016).
13. Howitt, M. R. et al. Tuft cells, taste-chemosensory cells, orchestrate parasite type 2 immunity in the gut. *Science* **351**, 1329–1333 (2016).
14. Gerbe, F. et al. Intestinal epithelial tuft cells initiate type 2 mucosal immunity to helminth parasites. *Nature* **529**, 226–230 (2016).
15. Hoover, B. et al. The intestinal tuft cell nanostructure in 3D. *Sci. Rep.* **7**, 1652 (2017).
16. Krasteva, G. et al. Cholinergic chemosensory cells in the trachea regulate breathing. *Proc. Natl Acad. Sci. USA* **108**, 9478–9483 (2011).
17. Deckmann, K. et al. Bitter triggers acetylcholine release from polymodal urethral chemosensory cells and bladder reflexes. *Proc. Natl Acad. Sci. USA* **111**, 8287–8292 (2014).
18. Panneck, A. R. et al. Cholinergic epithelial cell with chemosensory traits in murine thymic medulla. *Cell Tissue Res.* **358**, 737–748 (2014).
19. Soultanova, A. et al. Cholinergic chemosensory cells of the thymic medulla express the bitter receptor Tas2r131. *Int. Immunopharmacol.* **29**, 143–147 (2015).
20. Bezençon, C. et al. Murine intestinal cells expressing Trpm5 are mostly brush cells and express markers of neuronal and inflammatory cells. *J. Comp. Neurol.* **509**, 514–525 (2008).
21. Wong, G. T., Gannon, K. S. & Margolskee, R. F. Transduction of bitter and sweet taste by gustducin. *Nature* **381**, 796–800 (1996).
22. Zhang, Y. et al. Coding of sweet, bitter, and umami tastes: different receptor cells sharing similar signaling pathways. *Cell* **112**, 293–301 (2003).
23. Yano, M. et al. Aire controls the differentiation program of thymic epithelial cells in the medulla for the establishment of self-tolerance. *J. Exp. Med.* **205**, 2827–2838 (2008).
24. Rattay, K. et al. Homeodomain-interacting protein kinase 2, a novel autoimmune regulator interaction partner, modulates promiscuous gene expression in medullary thymic epithelial cells. *J. Immunol.* **194**, 921–928 (2015).
25. Lee, Y. J. et al. Tissue-specific distribution of iNKT cells impacts their cytokine response. *Immunity* **43**, 566–578 (2015).
26. Lee, Y. J., Holzapfel, K. L., Zhu, J., Jameson, S. C. & Hogquist, K. A. Steady-state production of IL-4 modulates immunity in mouse strains and is determined by lineage diversity of iNKT cells. *Nat. Immunol.* **14**, 1146–1154 (2013).
27. Matsumoto, I., Ohmoto, M., Narukawa, M., Yoshihara, Y. & Abe, K. Skn-1a (Pou2f3) specifies taste receptor cell lineage. *Nat. Neurosci.* **14**, 685–687 (2011).

Acknowledgements We thank A. Chan, M. Waterfield, L. Velten and the Anderson, Locksley and Steinmetz laboratories for helpful discussions; K. Wu, Y. Wang and E. Li for experimental support. Tetramers were from the NIH Tetramer Core Facility. *Pou2f3*$^{-/-}$ mice were from the DTCC-KOMP2 Consortium from The Canadian Mouse Mutant Repository. Biostatistics support was provided by the UCSF Functional Genomics Core. This work was supported by NIH grant R01 AI097457 (C.N.M., I.P., I.S.K., T.C.M. and M.S.A.); Larry Hillblom Foundation 2017-D-012-FEL (I.P.); NIH Medical Scientist Training Program grant T32 GM007618 to UCSF (I.S.K.); NSF GRFP DGE 1656518 (K.L.W.); NSF GRFP DGE 1656518 (A.R.R.); NIH grant R37 AI039560 (H.W. and K.A.H.); Damon Runyon Cancer Research Foundation DRG-2162-13 (J.v.M.); Howard Hughes Medical Institute (J.v.M. and R.M.L.); Sandler Asthma Basic Research Center (A.C.F., J.v.M. and R.M.L.); German Cancer Research Center (K.R.); European Research Council grant ERC-2012-AdG (B.K.); NIH grant U01 DK107383 (A.V.P. and M.S.A.); NIH grant P01 HG00020527 (L.M.S.); NIH grant R01 AI026918 (R.M.L.); NIH Diabetes Research Center grant P30 DK063720 (A.C.F., M.S.A. and Single Cell Analysis Center); NIH Shared Instrument Grant 1S10OD021822-01 (Single Cell Analysis Center).

Reviewer information *Nature* thanks H.-R. Rodewald and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions C.N.M. conceived the study, designed and performed experiments and analysed data. I.P. conceived and designed experiments with C.N.M., performed thymocyte flow cytometry and analysed data. J.v.M. designed and performed bulk tuft RNA-seq, provided Flare25 and *Trpm5*$^{-/-}$ mice and analysed data with R.M.L. K.L.W. and A.R.R. designed and performed single-cell RNA-seq and analysed data with L.M.S. H.W. designed and performed F_1 experiments and analysed data with K.A.H. K.R. designed and performed flow cytometry on thymic *Hipk2*-knockout mTECs and analysed data with B.K. I.S.K. and T.C.M. designed and performed mTEC iALT RNA-seq. J.L.P. designed and performed biostatistical analysis of RNA-seq data and analysed data with D.J.E. A.C.F. designed and performed quantitative immunofluorescence image analysis. A.V.P. designed and performed human mTEC experiments. E.J.W. designed and performed immunofluorescence experiments and analysed data. W.W.L. performed RLBA experiments. M.S.A. directed the study and wrote the manuscript with C.N.M. and I.P.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0345-2.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0345-2.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to M.S.A.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Mice. *Aire*$^{-/-}$, *Aire*^{creERT2}, *Aire*^{DTR}, iALT, Flare25, *Hipk2*$^{fl/fl}$, *Trpm5*$^{-/-}$, *Kn2* BALB/cByJ and *Foxn1*^{cre} mice have been described previously[4,12,24,26–30]. *Pou2f3*$^{-/-}$ (*Pou2f3*$^{tm1.1(KOMP)Vlcg}$, Project ID VG18280) sperm was obtained from the DTCC-KOMP2 Consortium and animals were rederived by the UCSF Rederivation Core. Flare25 and *Trpm5*$^{-/-}$ mice were provided by J.v.M. and R.M.L. (UCSF). C57BL/6 (JAX 000664), BALB/cByJ (JAX 001026), C57BL/6Cg-*Gt(Rosa)*$^{26Sortm14(CAG-tdTomato)Hze}$/J (JAX 007914), C57BL/6Cg-*Gt(Rosa)26Sor*$^{tm6(CAG-ZsGreen1)Hze}$/J (JAX 007906) mice were obtained from The Jackson Laboratory[31].

Most mice were maintained in the University of California San Francisco (UCSF) specific pathogen-free animal facility in accordance with the guidelines established by the Institutional Animal Care and Use Committee (IACUC) and Laboratory Animal Resource Center and all experimental procedures were approved by the Laboratory Animal Resource Center at UCSF. Some mice were maintained in the German Cancer Research Center (DKFZ) specific pathogen-free animal facility in accordance with the European Convention for the Protection of Vertebrate Animals used for Experimental and Other Specific Purposes and the German Legislation. Some mice were maintained at the University of Minnesota under specific-pathogen-free conditions in accordance with the guidelines established by the IACUC. Mice aged 4–8 weeks were used for all experiments unless otherwise specified in the text or figure legends. Mice were age-matched in figures displaying a single representative experiment and in pooled data. Mice were sex-matched when possible.

In vivo mouse treatments. Diphtheria toxin (Sigma-Aldrich) was administered to *Aire*^{DTR} mice at a dose of 25–50 ng g^{-1} every other day via intraperitoneal injections for a total of five injections for all experiments. For tamoxifen treatment of mice possessing conditional alleles, tamoxifen (Sigma-Aldrich) was dissolved in corn oil (Sigma-Aldrich) and 2-mg doses were administered by oral gavage with flexible plastic feeding tubes (Instech). For continuous treatment, animals were gavaged every other day until the experimental end point.

Mice were immunized with 100 μg of recombinant mouse IL-25 protein in PBS (R&D Systems) emulsified 1:1 v/v in CFA containing 4 mg ml^{-1} heat-inactivated *Mycobacterium tuberculosis* strain H37a (Chondrex). A total of 200 μl was injected subcutaneously, divided equally between both sides. Three weeks later, serum was collected for analysis and stored at −20 °C until radioligand-binding analysis.

Bulk RNA-seq sample preparation and analysis. Single-cell epithelial suspensions were isolated and stained as described below and then sorted using a FACS Aria III or Aria Fusion (BD Biosciences). Cells were sorted directly into DMEM (ThermoFisher) containing 10% FBS, pelleted and snap-frozen in liquid nitrogen before total RNA extraction. For lineage tracing, approximately 80,000 pre-*Aire*, 150,000 early-*Aire*, 150,000 late-*Aire* and 25,000 post-*Aire* cells were collected for each biological replicate. For Flare25 RFP^{+} small-intestinal and thymic tuft cells, approximately 35,000 small-intestinal RFP^{+} and 1,750 thymic RFP^{+} cells were collected for each biological replicate. Intact mRNA was isolated using the Dynabead mRNA Purification Kit for total RNA (ThermoFisher) and amplified cDNA was prepared using the NuGen Ovation RNA-Seq system V2 kit, according to the manufacturer's protocol (NuGen Technologies). Sequencing libraries were generated using the Nextera XT library preparation kit with multiplexing primers, according to manufacturer's protocol (Illumina). Library fragment-size distributions were assessed using the Bioanalyzer 2100 and the DNA high-sensitivity chip (Agilent Technologies). Library sequence quality was assessed by sequencing single-end 50-bp reads using the Illumina MiSeq platform and libraries were pooled for high-throughput sequencing on the Illumina HiSeq 4000 by using equal numbers of uniquely mapped reads (Illumina). Then, 12 samples per lane were multiplexed to ensure adequate depth of coverage. The analytic pipeline included demultiplexing raw sequencing results, trimming adaptor sequences, and aligning to the reference genome. Sequence alignment and splice junction estimation was performed using the STAR (https://code.google.com/p/rna-star) software program. For differential expression testing, the genomic alignments were restricted to those that map uniquely to the set of known Ensembl IDs (including all protein-coding mRNAs and other coding and noncoding RNAs). STAR-aggregated mappings on a per-gene basis were used as raw input for normalization by DESeq2 software.

Single-cell RNA-seq and computational analysis. Single-cell epithelial suspensions were isolated and stained as described below and sorted using a FACS Aria Fusion (BD Biosciences). Single cells were sorted into individual wells of a 96-well plate (Bio-Rad) containing 5 μl of lysis solution and sequencing libraries were prepared using the smart-seq2 method as reported with the following modifications[32]. Batch one, 0.1 μl of a 1:1,000,000 dilution of ERCC Spike-In Mix (Life Technologies) in RNase-free water was included in the reverse transcription master mix. Batch 2, 1 μl of a 1:1,000,000 dilution of ERCC Spike-In Mix (ThermoFisher) in RNase-free water was included in a total volume of 5 μl lysis buffer. For both batches, 21 cycles of initial PCR amplification were performed and a ratio of 0.56:1.0 (beads:total PCR volume; instead of 1.0:1.0) of Ampure XP beads (Beckman Coulter) was used for the first PCR purification to minimize primer dimer carryover. After the first PCR amplification, cDNA libraries were screened via quantitative PCR (we used a 1:10 dilution of purified cDNA libraries for quantitative PCR) for expression of a mouse housekeeping gene (*Ubc*), and the distribution of library size was checked on a Bioanalyzer instrument (Agilent). Only cDNA libraries that passed both quality controls were processed further. Subsequently, 50 pg of cDNA was used for the 'tagmentation' (transposase-based fragmentation) reaction and 12 cycles were applied for the final enrichment PCR. The final purification step was performed with a ratio of 0.6:1.0 (as above) of Ampure XP beads (Beckman Coulter).

In batch one, 96 cells were initially 'multiplexed' with 48 samples per Illumina HiSeq 2500 lane. Then, 31 additional cells were multiplexed with 96 samples per Illumina HiSeq 2500 lane and 58-bp single-end sequencing was performed. We observed no batch effect between these two sets. In batch two, 92 Flare25 cells and 96 Flare25 *Aire*$^{-/-}$ cells were multiplexed with 96 samples per Illumina HiSeq 2500 lane. A negative control and a 50-cell positive control were sequenced from each sample. Flare25 and Flare25 *Aire*$^{-/-}$ cells were split evenly between sequencing lanes and 85-bp single-end sequencing was performed. HiSeq sequencing lanes typically yielded between ~150×10^6 and ~200×10^6 reads for both batches.

Raw data reads were aligned to the GRCm38 build of the *Mus musculus* genome using the STAR alignment software[33]. Cells were required to have a minimum of 40% total reads mapping to the genome, which all of our cells met. Reads uniquely aligned to mRNA genes were counted with featureCounts, a part of the Subread software package[34]. Genes present in a minimum of five cells and with at least 10 reads were retained in the analysis. Cells expressing fewer than 750 genes with 10 reads were removed from the analysis. All cells were required to have a minimum of 100,000 total reads mapping to genes, and less than 10% of reads mapping to mitochondrial genes. This resulted in 197 Flare25 cells and 95 Flare25 *Aire*$^{-/-}$ cells for further downstream analysis.

Data were normalized using the indeXplorer pipeline (https://git.embl.de/velten/indeXplorer)[35]. In brief, data were fitted to error models using the SCDE software to model dropout events and overdispersion of genes and the posterior odds ratio was used as a normalized measure of gene-expression values[35,36].

The effect of mitochondrial read counts were regressed out before all further downstream analysis. Principle component analyses were performed on the total set of cells, including both Flare25 and Flare25 *Aire*$^{-/-}$ cells, using the 'prcomp' function from the 'stats' R package. Heat maps were generated by calculating the mean centred expression values for each gene across all cells and clustered within genotype (independently for Flare25 and Flare25 *Aire*$^{-/-}$) using hierarchical clustering with Ward linkage ('hclust' from the 'stats' R package). Genes were ordered by hierarchical clustering with Ward linkage across all cells. Histograms were generated by counting the number of cells with the associated number of *Tas2r* genes present or not present (defined as normalized expression value greater than 0). Correlation values between *Tas2r* gene pairs were calculated using Spearman's rank correlation. Statistically significant correlations were established by constructing an empirical cumulative distribution function for all pairwise correlation values for all genes. Pairwise correlation values falling within the high correlation cutoff of 0.05 were determined significantly high by an empirical P value of $P < 0.05$.

Immunofluorescence staining and imaging. For immunofluorescence, tissues were fixed in 2% paraformaldehyde (Pierce) in PBS for 2 h at 4 °C followed by overnight incubation in 30% (w/v) sucrose (Sigma-Aldrich) in PBS. Tissues were embedded in Optimal Cutting Temperature Compound (Tissue-Tek) and stored at −80 °C before sectioning (50–200 μm) on a cryostat (Leica). Thin sections were dried on Superfrost Plus (Fisher Scientific) slides and semi-thick (200 μm) sections were moved directly to 0.3% Triton X-100 (Sigma-Aldrich), 0.2% BSA (Sigma-Aldrich), 0.1% sodium azide (Sigma-Aldrich) in PBS (Immunomix). Slides were stained in a humidified chamber and semi-thick sections were stained in 24-well plates with one section per well. Slides were briefly rehydrated in PBS before permeabilization in Immunomix for 1 h at room temperature followed by blocking with BlockAid (ThermoFisher), primary antibody staining at room temperature, and, when necessary, secondary antibody staining at room temperature for 1 h. Semi-thick sections were permeabilized in Immunomix with shaking at room temperature overnight followed by blocking with BlockAid at room temperature for 2 h, primary antibody staining at room temperature for 2 h, and, when necessary, secondary antibody staining at room temperature for 2 h. Semi-thick sections were then moved to Superfrost Plus slides and all sections were mounted with ProLong Diamond Antifade Mountant (ThermoFisher). Images were acquired on a Leica SP5 (Leica) laser scanning confocal microscope. The following antibodies were used in this study: DCLK1 (Abcam polyclonal ab31704), KRT10 (EP1607IHCY), KRT5 (EP1601Y), KRT8 (EP1628Y), tdTomato (632496). Antibodies were purchased from Abcam or Clontech.

Human thymic tissue. Thymus samples were collected from paediatric patients undergoing corrective cardiothoracic surgery. All experiments were performed in accordance with protocols approved by the UCSF Human Research Protection

© 2018 Springer Nature Limited. All rights reserved.

Program Institutional Review Board and written informed consent was obtained before sample collection.

Immunohistochemistry. Tissue was fixed in 4% paraformaldehyde (ThermoFisher) for 3 h at 4 °C, washed with PBS, embedded in paraffin and sectioned. Antigen retrieval was performed on rehydrated tissue by boiling sections in antigen retrieval Citra Solution (Biogenex). Sections were blocked for 30 min at room temperature using CAS-Block (ThermoFisher) with 0.2% Triton X-100 (Sigma-Aldrich), followed by incubation with anti-DCLK1 primary antibody (Abcam, ab31704) overnight at 4 °C. Staining with biotinylated secondary antibody was performed for 1 h at room temperature. Slides were developed using an ABC kit (Vector labs) and DAB kit (Vector labs) and counterstained with haematoxylin.

Image processing and quantitative image analysis. The spatial distribution of tuft cells was evaluated using the Spatstat R library. Raw Leica image files of three large scans ($3{,}670 \times 3{,}670 \times 77$ μm) were imported into Imaris 8.3.1 (Bitplane). Intensity-based three-dimensional segmentation was then applied to 12 identically sized regions of interest ($586 \times 272 \times 77$ μm) using identical thresholding parameters to generate surface coordinates of KRT10 bodies and centre of intensity coordinates of tuft cells. These coordinates were flattened to two dimensions (xy) and duplicates were removed. Two-dimensional coordinates were imported into R for spatial analysis and the relationship between surfaces and centre of intensity coordinates was calculated using the multi-type pair correlation function (PCF) in R, pcfcross (http://spatstat.org/). The pair-correlation function is defined as $g = K'(r)/2\pi r$, where $K'(r)$ is the first derivative of Ripley's K-function. Because PCF analysis is sensitive to spatial inhomogeneities within a field, the 12 ROIs were selected to avoid inhomogeneities at the field margins.

To measure the range of cluster sizes of the tuft cell aggregates, raw Leica images of fixed thymic slices were analysed in the Imaris v.8.3.1 software package (Bitplane). Intensity-based three-dimensional segmentation was applied to the image voxels and volumes of the tuft cells were measured. Surface volumes <5 times the mean single-cell volume (1×10^4 μm^3) were shown in green and surface volumes >5 (1×10^4 μm^3) were pseudocoloured based on aggregate size.

Single-cell tissue preparation. Mouse thymi were isolated, cleaned of fat and transferred to DMEM (UCSF Cell Culture Facility) containing 2% FBS (Atlanta Biologics) on ice. Single-cell suspensions of thymocytes were prepared by mashing through a 40-μm filter (Falcon). For mTEC preparations, thymi were minced with a razor blade. Up to five thymi were pooled into a single digestion. Tissue pieces were moved with a glass Pasteur pipette to 15-ml tubes and vortexed briefly in 10 ml of medium. Fragments were allowed to settle before removing the medium and replacing it with 4 ml of digestion medium containing 2% FBS, 100 μg ml^{-1} DNase I (Roche) and 100 μg ml^{-1} liberase TM (Sigma-Aldrich) in DMEM. Tubes were moved to a 37 °C water bath and fragments were triturated through a glass Pasteur pipette at 0 min and 6 min to mechanically aid digestion. At 12 min, tubes were spun briefly to pellet undigested fragments and the supernatant was moved to 20 ml of 0.5% BSA (Sigma-Aldrich), 2 mM EDTA (TekNova) in PBS (MACS buffer) on ice to stop the enzymatic digestion. This was repeated twice for a total of three 12-min digestion cycles, or until there were no remaining tissue fragments. The single-cell suspension was then pelleted and washed once in MACS buffer. Density-gradient centrifugation using a three-layer Percoll gradient (GE Healthcare) with specific gravities of 1.115, 1.065 and 1.0 was used to enrich for stromal cells. Cells isolated from the Percoll-light fraction, between the 1.065 and 1.0 layers, were then resuspended in MACS buffer and counted.

Human thymus was transferred to RPMI (ThermoFisher) and cut into small pieces using scissors. Tissue pieces were mashed gently using the back of a sterile syringe to extract most of the thymocytes. The supernatant was removed and replaced with fresh RPMI. Tissue pieces were moved to a 15-ml tube using a 5-ml pipette. Fragments were allowed to settle before discarding the medium and replacing it with 5 ml of digestion medium containing 100 μg ml^{-1} DNase I (Roche) and 100 μg ml^{-1} Liberase TM (Sigma-Aldrich) in RPMI. Tubes were moved to a 37 °C water bath and fragments were triturated through a 5-ml pipette at 0 min and 6 min to mechanically aid digestion. At 12 min, tubes were spun briefly to pellet undigested fragments and the supernatant was discarded. Fresh digestion medium was added to remaining fragments and the digestion was repeated using a glass Pasteur pipette for trituration. Supernatant from this second round of digestion was also discarded. A third round of enzymatic digestion was performed using 5 ml of digestion medium supplemented with trypsin–EDTA (ThermoFisher) for a final concentration of 0.05%. Remaining thymic fragments were digested for an additional hour with trituration every 15 min. The cells were moved to cold MACS buffer to stop the enzymatic digestion. TECs were enriched and prepared for flow cytometry using the same protocol as mouse mTECs with the following modifications: blocking was done with human Fc Receptor Binding Inhibitor Monoclonal Antibody (eBioscience) and human-specific antibodies against EPCAM, CD45 and HLA-DR were used for surface staining.

Small intestines were flushed with PBS, opened and rinsed with PBS to remove luminal contents. Subsequently, 2.5–5-cm-long segments of jejunum were incubated with rocking for 20 min at 37 °C in 5 ml PBS containing 2.5 mM EDTA (Sigma-Aldrich), 0.75 mM dithiothreitol (Sigma-Aldrich), and 10 μg ml^{-1} DNaseI (Sigma-Aldrich). Tissues were shaken vigorously for 30 s and released cells were incubated with rocking for 10 min at 37 °C in 5 ml HBSS (Ca^{2+}/Mg^{2+} free) containing 1.0 U ml^{-1} dispase (Gibco) and 10 μg ml^{-1} DNaseI. Digested cells were passed through a 70-μm filter and washed once before staining for flow cytometry.

Flow cytometry and antibodies. For surface staining, single-cell suspensions were prepared as described and incubated with Live/Dead Fixable Blue Dead Cell Stain (ThermoFisher) in PBS for 20 min at 4 °C followed by blocking with anti-mouse CD16/CD32 (24G2) (UCSF Hybridoma Core Facility) and 5% normal rat serum for 20 min at 4 °C. Cells were then washed in FACS buffer and stained for surface markers for 30–45 min at 4 °C. For CD1d tetramer staining, thymocytes were incubated with BV421-conjugated PBS-57-loaded CD1d tetramer (NIH Tetramer Core) for 1 h at room temperature. DCLK intracellular staining of mTECs was performed as described previously[12]. For staining of intracellular transcription factors, thymocytes were fixed and permeabilized with FoxP3 staining buffer set (eBioscience) according to the manufacturer's instructions. Flow cytometry data were collected on a LSRII Flow Cytometer (BD Biosciences) housed within the UCSF Single Cell Analysis Center, and analysed using FlowJo 10.3 software (TreeStar Software). The following antibodies were used in this study: DCLK1 (Abcam polyclonal ab31704), DCLK1 (EPR6085) Ly51 (6C3), CD11c (N418), CD45 (30-F11), EpCAM (G8.8), AIRE (5H12), I-A^b (25-9-17), I-A^d (39-10-8), TCRβ (H57-597), CD4 (GK1.5), CD8 (53-6.7), FoxP3 (MF-14), CD25 (3C7), PLZF (Mags21F7), T-bet (4B10), RORγ (Q31-378), EOMES (Dan11mag), CD2 (S5.2) human EPCAM (HEA-125), human CD45 (HI30), human HLA-DR (L243). Antibodies were purchased from Abcam, BioLegend, BD Biosciences, eBioscience or Miltenyi.

Quantitative RT–PCR. Single-cell epithelial suspensions were isolated and stained as described above and then sorted using a FACS Aria Fusion (BD Biosciences). Cells were sorted directly into TRIzol LS (ThermoFisher Scientific) and the aqueous phase was isolated using Phasemaker Tubes (ThermoFisher) and RNA was isolated using the Micro Plus RNeasy kit (Qiagen) and reverse transcribed using the SuperScript IV Vilo Master Mix with ezDNase (ThermoFisher). The resulting cDNA was used as template for quantitative PCR with Taqman Gene Expression Assays on a 7500 Fast Real-Time PCR System (Applied Biosystems). Transcripts were normalized to *Actb* (Mm00607939_s1) expression. *Aire* (Mm00477461_m1), *Dclk1* (Mm00444950_m1), *Epcam* (Mm00493214_m1), *Il25* (Mm00499822_m1) and *Trpm5* (Mm01129032_m1) primers were used.

Radioligand-binding assay. Full-length mouse *Il25* cDNA (MMM1013-211691449, Dharmacon) was transcribed, in vitro translated and biosynthetically labelled with ^{35}S-methionine using the TNT Quick Coupled Transcription/Translation System (Promega). Radiolabelled protein was immunoprecipitated with serum samples loaded in triplicate in 96-well PVDF filtration plates (Millipore). Radioactivity retained in the filter was measured using a liquid scintillation counter (1450 MicroBeta Trilux; PerkinElmer). A polyclonal goat anti-mouse IL-25 antibody (R&D Systems, af1399) was used as a positive control. IL-25 autoantibody indices for samples were calculated as: (c.p.m. of unknown – c.p.m. of negative standard)/(c.p.m. of positive standard – c.p.m. of negative standard) × 100. The limit of detection was defined as the mean of autoantibody indices for all wild-type samples plus three standard deviations.

Statistical analysis. All experiments were performed using randomly assigned mice without investigator blinding. No data were excluded. Statistical significance between two groups was calculated using an unpaired, parametric, two-tailed Student's t-test. Where more than two groups were plotted together, statistical significance was calculated either by using one-way, non-parametic ANOVA (Kruskal–Wallis test) or using two-way non-parametric ANOVA with multiple comparisons. Experimental groups included a minimum of three biological replicates. Intragroup variation was not assessed. All statistical analysis was performed using Prism 7 (GraphPad Software), except where stated above. Figures display mean ± s.d. A P value of less than 0.05 was considered statistically significant. No statistical methods were used to predetermine sample size.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

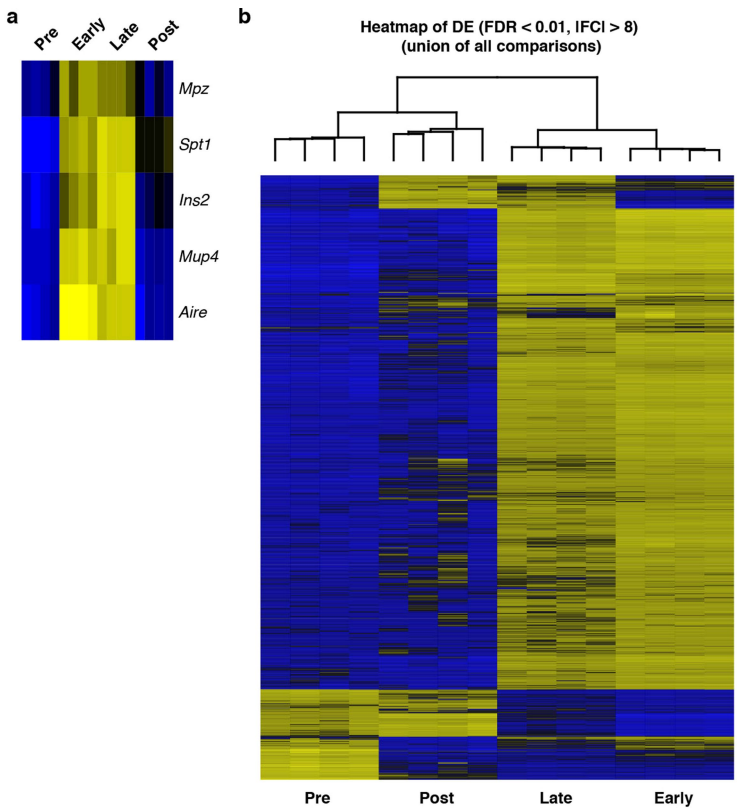
Code availability. Code used to align, perform quality control and analyse single-cell RNA-seq data can be found at https://github.com/mTEC-pipelines/tufts-pipelines. Bulk RNA-seq and quantitative image analysis pipelines are described and code is available upon request or through publicly available GitHub repositories.

Data availability. RNA-seq data that support the findings of this study have been deposited in the GEO database under accession numbers GSE114651, GSE114668 and GSE114895. Source Data for Figs. 1e, 2c, 3e, f, h, 4c–f and Extended Data Figs. 5d, f–i, 9b–d, 10b–e, g are provided in the online version of the paper. All other data that support the findings of this study are available from the corresponding author upon reasonable request.

© 2018 Springer Nature Limited. All rights reserved.

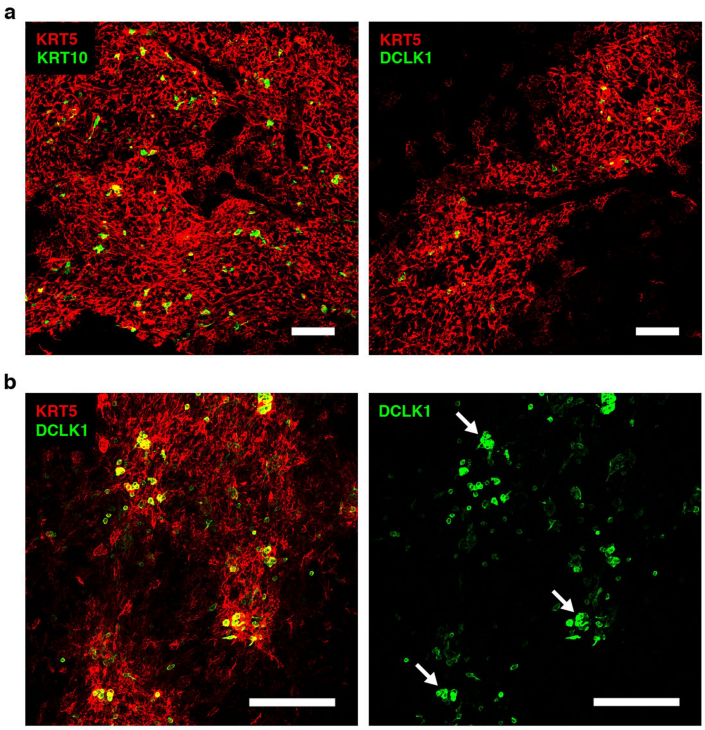
28. Anderson, M. S. et al. Projection of an immunological self shadow within the thymus by the Aire protein. *Science* **298**, 1395–1401 (2002).
29. Damak, S. et al. Trpm5 null mice respond to bitter, sweet, and umami compounds. *Chem. Senses* **31**, 253–264 (2006).
30. Gordon, J. et al. Specific expression of lacZ and cre recombinase in fetal thymic epithelial cells by multiplex gene targeting at the Foxn1 locus. *BMC Dev. Biol.* **7**, 69 (2007).
31. Madisen, L. et al. A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. *Nat. Neurosci.* **13**, 133–140 (2010).
32. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
33. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
34. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
35. Velten, L. et al. Human haematopoietic stem cell lineage commitment is a continuous process. *Nat. Cell Biol.* **19**, 271–281 (2017).
36. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).

© 2018 Springer Nature Limited. All rights reserved.



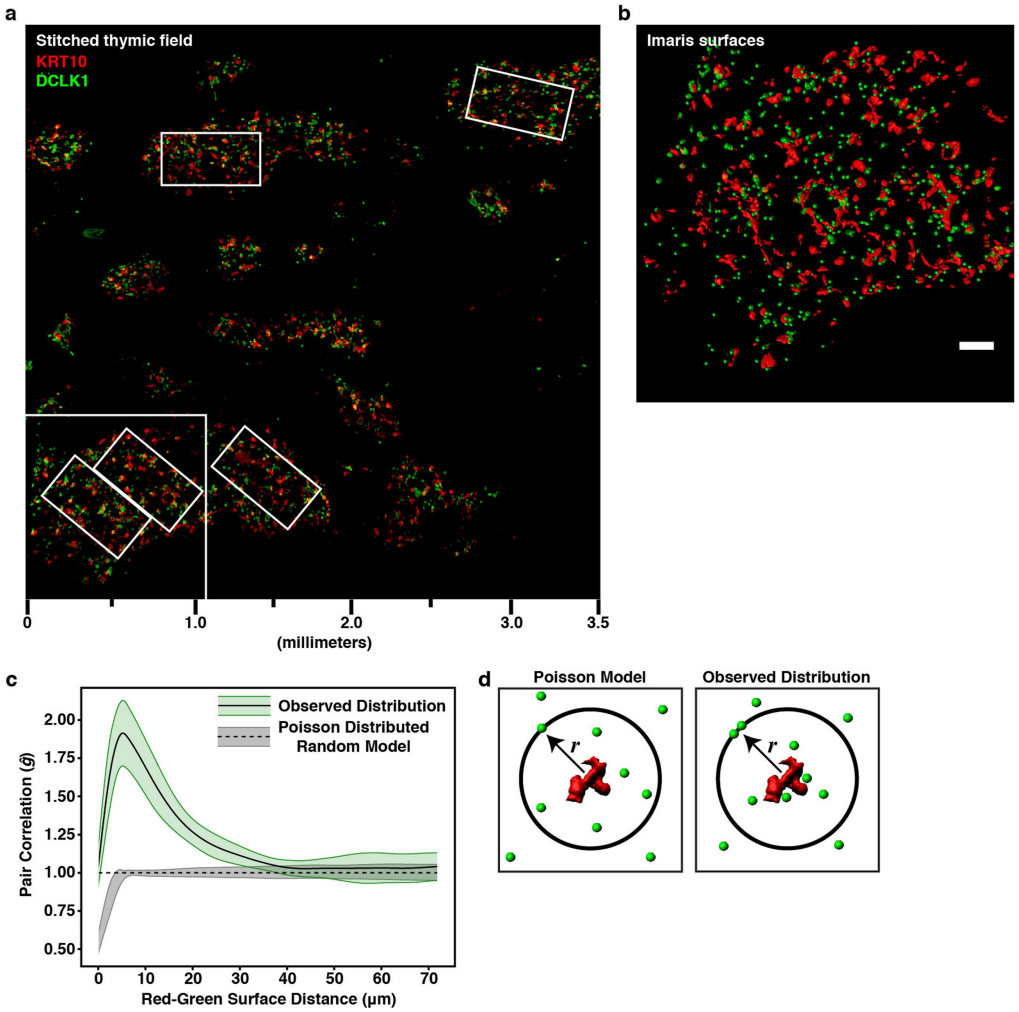
Extended Data Fig. 1 | RNA-seq of FACS-sorted iALT mTEC subsets. **a**, Heat map of *Aire* and representative *Aire*-dependent TSAs in pre-, early-, late- and post-*Aire* populations as defined in Fig. 1a. **b**, Heat map of differentially expressed genes with a FDR < 0.01 and |fold change| > 8. Columns are organized by unsupervised hierarchical clustering (union of all comparisons) with a dendrogram representing similarity between clustered columns. Note that the early- and late-*Aire*-expressing populations are grouped, as expected.

© 2018 Springer Nature Limited. All rights reserved.



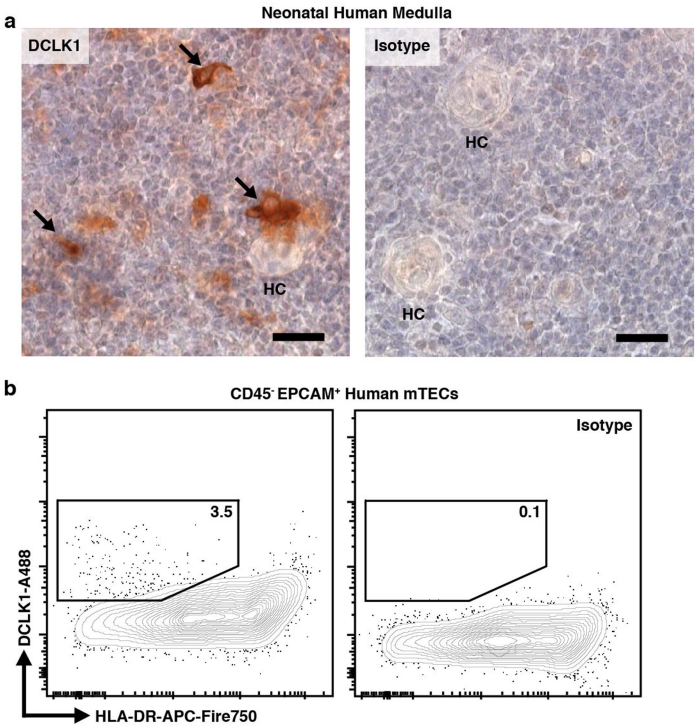
Extended Data Fig. 2 | Medullary localization of DCLK1^{bright} cells and KRT10 bodies. Representative confocal maximum projections (25 μm) of immunofluorescence staining of thymic slices at low magnification. **a**, KRT5 (red) and KRT10 (green) (left) and KRT5 (red) and DCLK1 (green) (right) showing medullary localization of KRT10 and DCLK1. **b**, Left, KRT5 (red) and DCKL1 (green). Right, DCLK1 alone. Region was selected for presence of multicellular DCLK1^{bright} cell clumps as indicated by white arrows. Scale bars, 100 μm. $n = 3$ mice; images are representative of three independent experiments.

© 2018 Springer Nature Limited. All rights reserved.



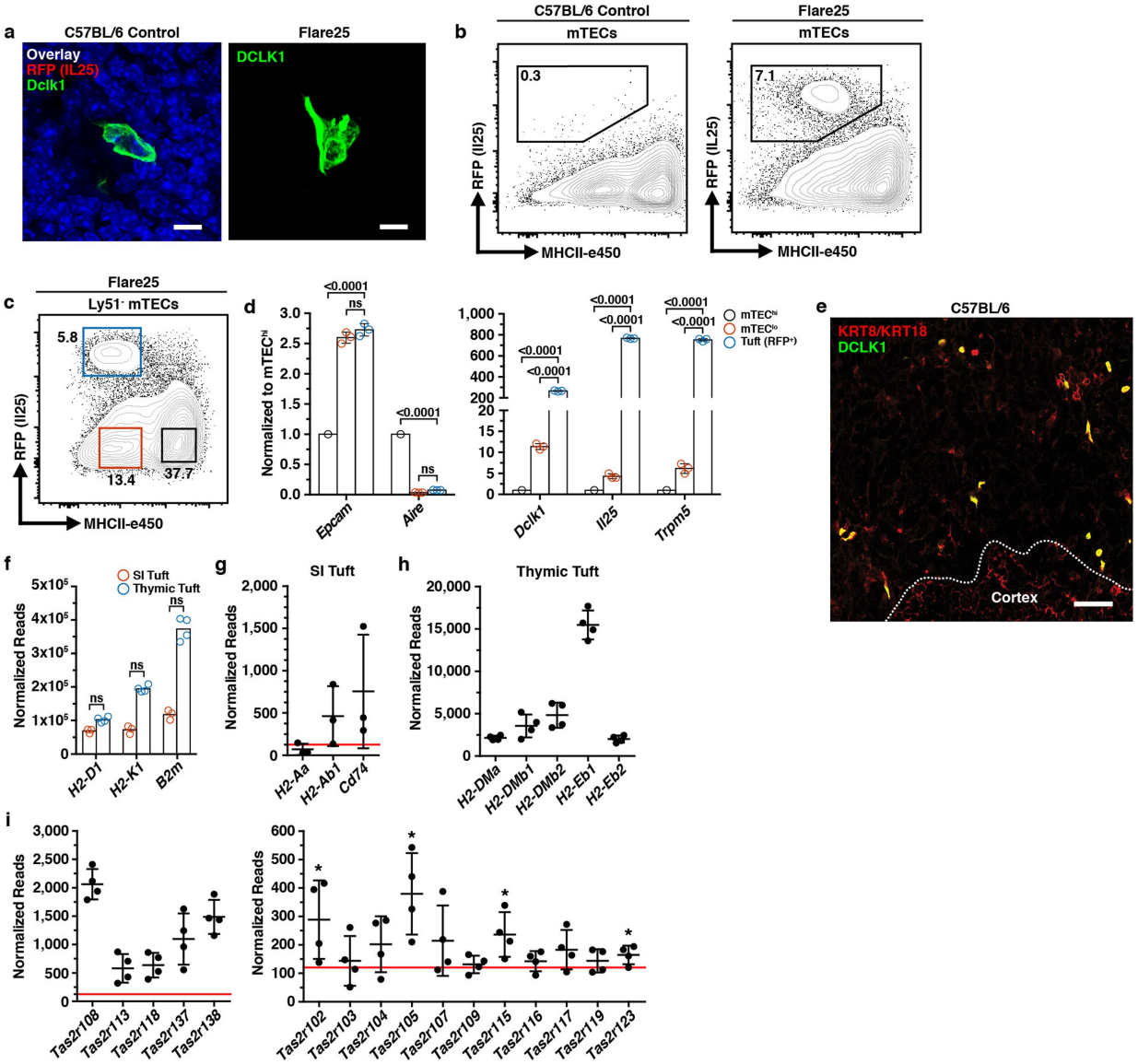
Extended Data Fig. 3 | Non-random distribution of medullary DCLK1^{bright} cells. a, Stitched confocal maximum projection (77 μm) of immunofluorescence staining of semi-thick (200 μm) thymic slice from a C57BL/6 mouse at low magnification. KRT10 (red) and DCLK1 (green). Field, 3,670 × 3,670 μm. Small rectangles indicate volumes selected for quantitative image analysis from this slice. **b**, Expanded area of image from **a** indicated in lower left corner with KRT10 signal converted into surfaces and DLCK1 signal converted to centre of intensity coordinates. Scale bar, 100 μm. **c**, Pair correlation function (PCF) analysis of 12 identically sized regions of interest (586 × 272 × 77 μm) from three different stitched thymic slices ($n = 3$ thymic slices). Results are presented as a histogram for a range of distances from KRT10 surfaces and are interpreted as follows: $g = 1$ indicates a spatial distribution that follows a random Poisson distribution, illustrated by a dashed line and grey envelope; $g < 1$ indicates regularity in the distribution; $g > 1$ indicates clustering. **d**, Schematic representation of the PCF analysis. The PCF counts the number of objects from a surface at a radial distance (r) and compares this number to the expected number of events for a random Poisson-distributed population at this distance.

© 2018 Springer Nature Limited. All rights reserved.



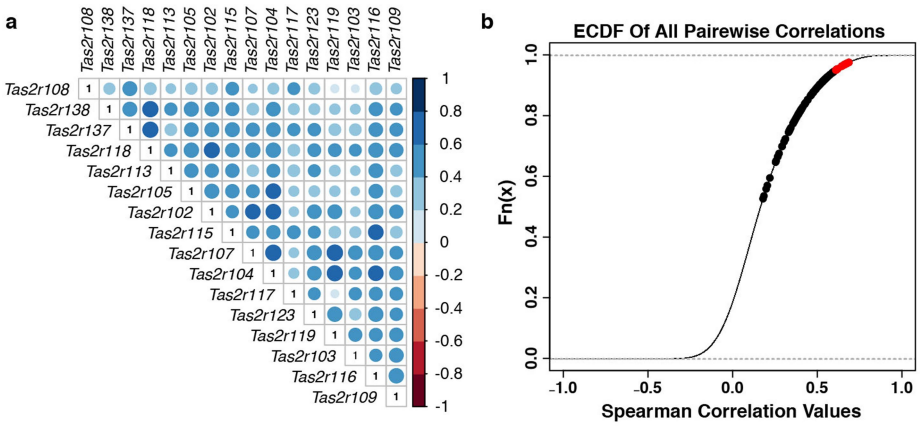
Extended Data Fig. 4 | Human thymus contains medullary DCLK1^{bright} cells. a, Representative immunohistochemistry of neonatal (21-day-old) human thymus stained for DCLK1 or isotype control. Arrows point to individual DCLK1^{bright} cells or clusters. HC, Hassall's corpuscle. Scale bars, 25 μm. **b**, Flow cytometry plot from enzymatically digested prenatal human thymus (22.3 gestational weeks) gated on CD45^{-}EPCAM^{+} TECs showing intracellular DCLK1 or isotype control. Note that human DCLK1^{+} events are approximately 3.5% of prenatal TECs. Data are representative of two independent experiments.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 5 | Characterization of thymic tuft cells. **a**, DCLK1 and RFP (IL-25) in the thymus of C57BL/6 control and Flare25 mice. Confocal maximal projection. Scale bar, 5 μm. $n = 3$ mice, two independent experiments. **b**, RFP (IL-25) in ($CD11c^{-}CD45^{-}EPCAM^{+}$) mTECs from C57BL/6 control and Flare25 mice. $n = 5$ mice; three independent experiments. **c**, Gating strategy for FACS sorting of $CD11c^{-}CD45^{-}EpCAM^{+}Ly51^{-}$ mTECs from Flare25-reporter mice for qPCR analysis of $mTEC^{high}$ ($RFP^{-}MHC\text{-}II^{high}$), $mTEC^{low}$ ($RFP^{+}MHC\text{-}II^{low}$), and thymic tuft ($RFP^{+}$) populations. **d**, qPCR analysis of expression of indicated genes of interest on populations sorted in **c** normalized to $mTEC^{high}$. Data are mean ± s.d. $n = 3$ mice. Two-way non-parametric ANOVA with multiple comparisons correction. **e**, Representative confocal maximum projection (10 μm) stained for KRT8/KRT18 (red) and DCLK1 (green). Scale bar, 50 μm. $n = 3$ mice, two independent experiments. **f**–**i**, Expression levels (normalized reads from Fig. 2b) from bulk RNA sequencing of small-intestinal (SI; $n = 3$ mice) and thymic tuft ($n = 4$ mice) cells. **f**, Expression of major MHC-I genes and *B2m*. FDR > 0.1. **g**, Expression of major MHC-II genes and *Cd74* in small-intestinal tuft cells. **h**, Expression of minor MHC-II genes in thymic tuft cells. **i**, Expression of *Tas2r* family members in thymic tuft cells. **g**–**i**, Data are mean ± s.d. **g**, **i**, The red line corresponds to a cut-off of 5 reads per million; data for which the mean ± s.d. fall above this cut-off are indicated with an asterisk.

© 2018 Springer Nature Limited. All rights reserved.



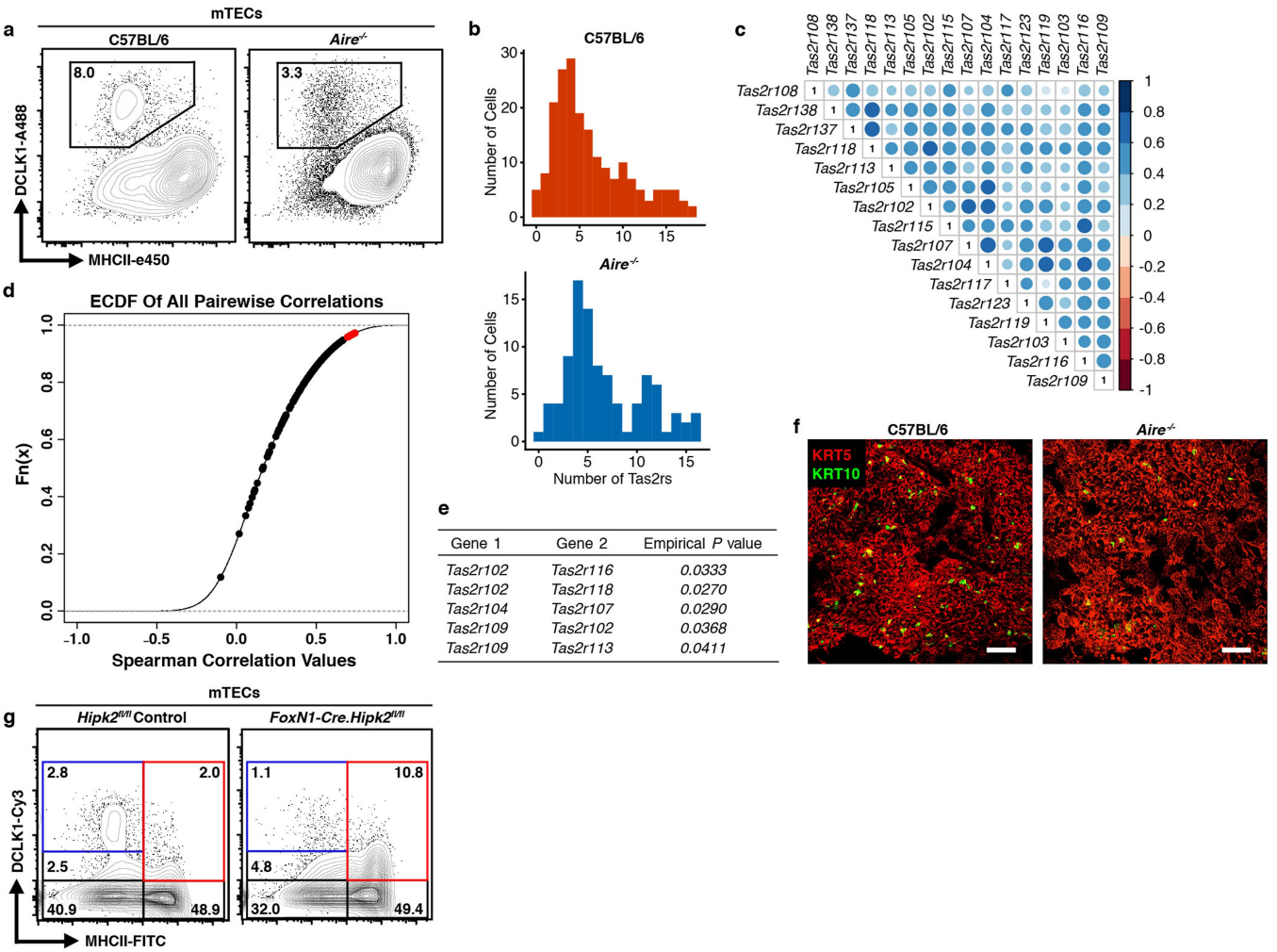
c

Significant Taste Receptor Pairwise Correlations

Gene 1	Gene 2	Empirical *P* value
Tas2r104	*Tas2r116*	0.0470
Tas2r115	*Tas2r116*	0.0448
Tas2r137	*Tas2r118*	0.0309
Tas2r102	*Tas2r118*	0.0356
Tas2r138	*Tas2r118*	0.0283
Tas2r107	*Tas2r119*	0.0231
Tas2r104	*Tas2r119*	0.0308
Tas2r104	*Tas2r105*	0.0447
Tas2r102	*Tas2r107*	0.0484
Tas2r104	*Tas2r107*	0.0264
Tas2r104	*Tas2r102*	0.0457

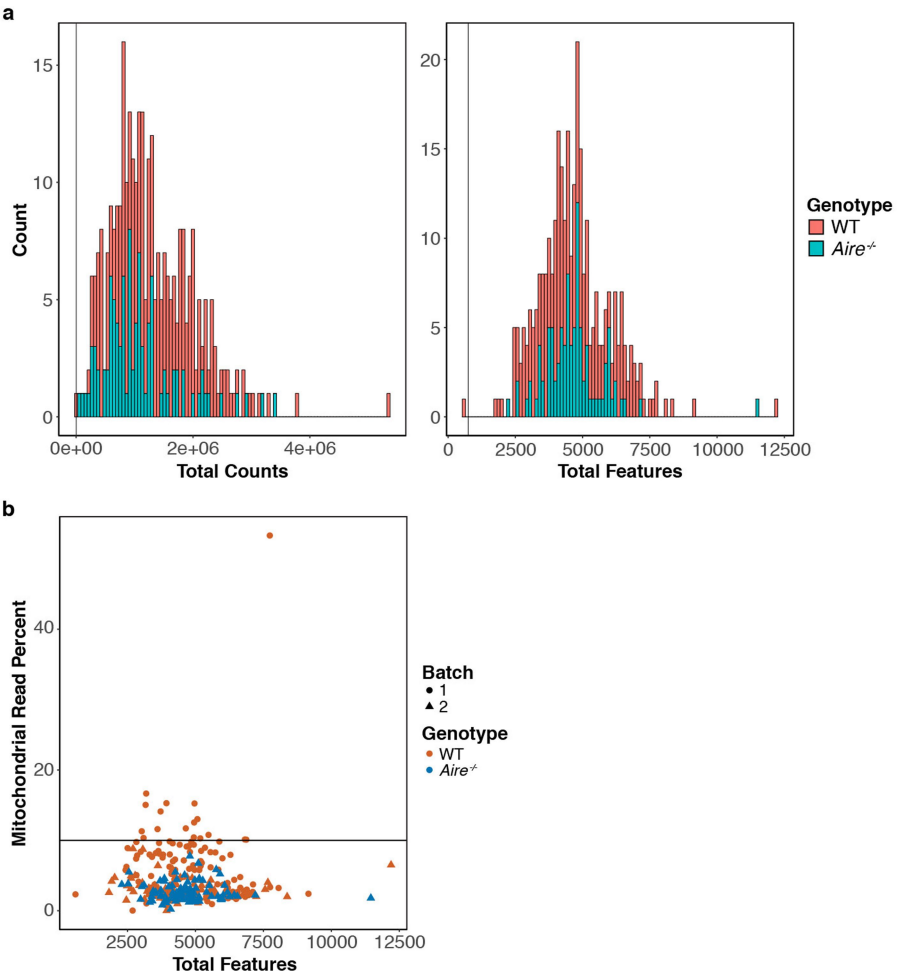
Extended Data Fig. 6 | *Tas2r* expression in single thymic tuft cells is not stochastic. a, Correlation plot of *Tas2r* genes in single Flare25 tuft cells as analysed in Fig. 2f. Colour and circle size represent pairwise correlation value. **b**, Empirical cumulative distribution function (ECDF) plot of pairwise gene expression correlation of single Flare25 tuft cells. Red points represent *Tas2r* genes that were significantly highly correlated compared to the background gene set ($P < 0.05$). **c**, List of correlated *Tas2r* gene pairs from **b** and their corresponding empirical *P* values.

© 2018 Springer Nature Limited. All rights reserved.



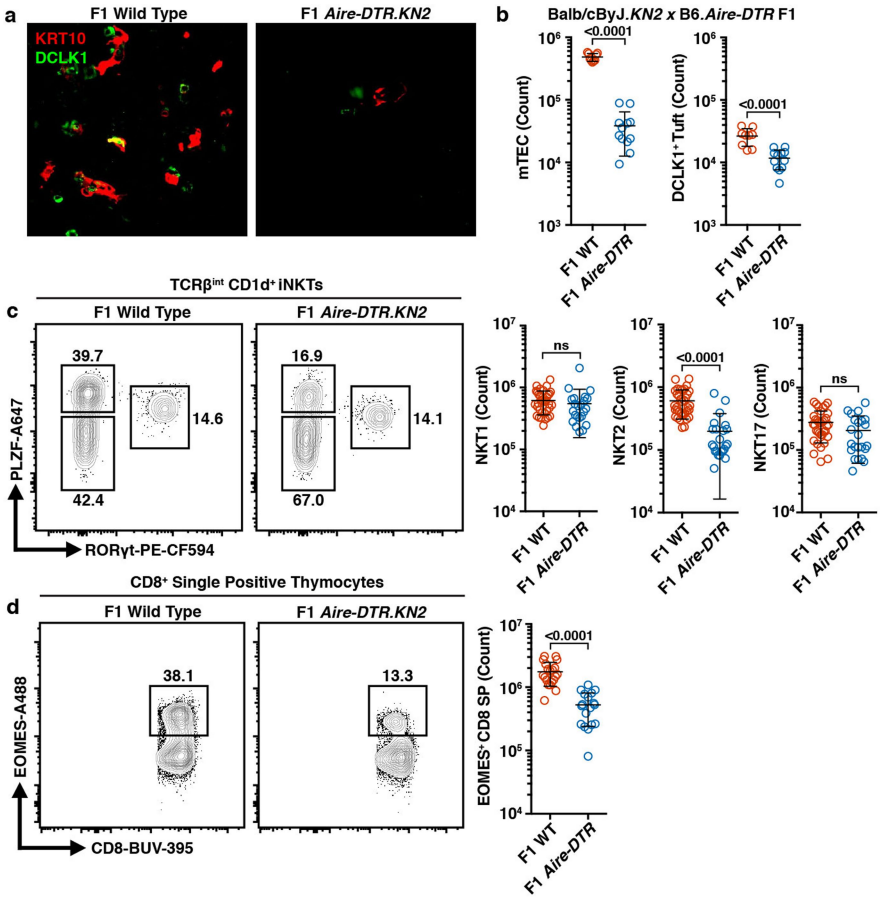
Extended Data Fig. 7 | Analysis of *Tas2r* expression in single Flare25 and Flare25;*Aire*$^{-/-}$ thymic tuft cells. a, Intracellular DCLK1 in C57BL/6 control and *Aire*$^{-/-}$ mTECs (CD11c^{-}CD45^{-}EPCAM^{+}). $n = 5$ mice; three independent experiments. **b**, Histogram showing the number of *Tas2r* family members expressed in each individual thymic tuft cell. Note that the distribution is similar between wild-type and *Aire*$^{-/-}$ cells. **c**, Correlation plot of *Tas2r* genes in single Flare25;*Aire*$^{-/-}$ tuft cells as analysed in Fig. 3c. Colour and circle size represent pairwise correlation values. **d**, Empirical cumulative distribution function plot of pairwise gene expression correlation of single Flare25;*Aire*$^{-/-}$ tuft cells. Red points represent *Tas2r* genes that were significantly highly correlated compared to the background gene set ($P < 0.05$). **e**, List of correlated *Tas2r* gene pairs from **d** and their corresponding empirical *P* values. **f**, Representative confocal maximum projections (25 μm) of immunofluorescence analysis of C57BL/6 control and *Aire*$^{-/-}$ thymi stained for KRT5 (red) and KRT10 (green). Scale bars, 100 μm. $n = 3$ mice; two independent experiments. **f**, Analysis of DCLK1^{+} cells in *Hipk2*$^{fl/fl}$ controls or *Foxn1*^{cre}*Hipk2*$^{fl/fl}$ (mTEC *Hipk2*$^{-/-}$) thymus gated on CD45^{-}CDR1^{-}EPCAM^{+} mTECs as quantified in Fig. 3h. $n = 6$ mice; three independent experiments.

© 2018 Springer Nature Limited. All rights reserved.



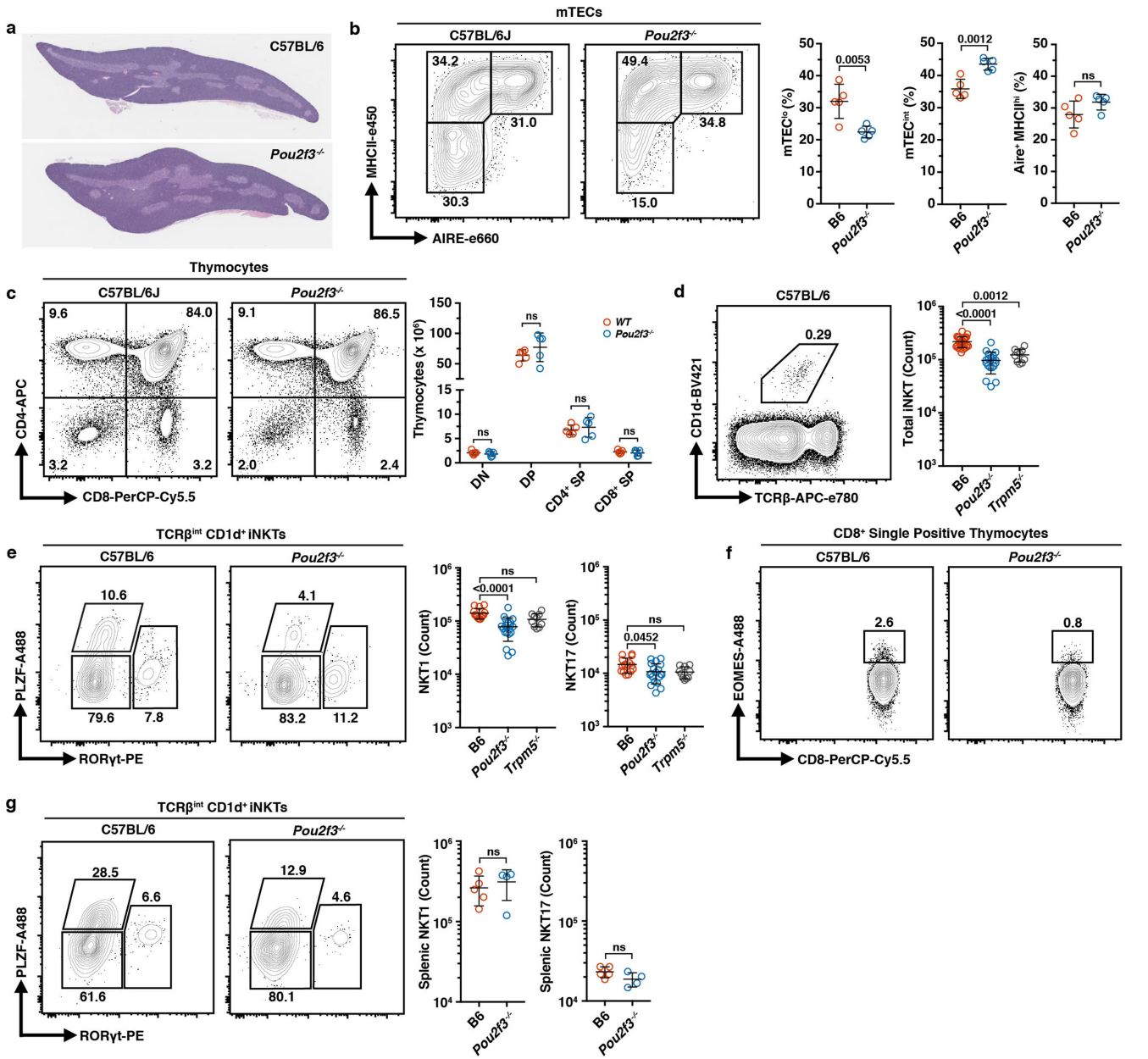
Extended Data Fig. 8 | Single-cell RNA-seq data quality and inclusion criteria. a, Histograms of total read counts (left) or features (right) for each single cell. Red represents Flare25 thymic tuft cells and blue represents Flare25;*Aire*$^{-/-}$ thymic tuft cells. Cells with fewer than 100,000 reads or 750 detected features were discarded from further analysis. **b**, Scatter plot of total features versus mitochondrial read percentage. Red represents Flare25 thymic tuft cells and blue represents Flare25;*Aire*$^{-/-}$ thymic tuft cells. Triangles and circles denote cells from two separate sorts. Cells with more than 10% of their reads mapping to mitochondrial genes were discarded from further analysis as indicated by the black line.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | Diphtheria toxin ablation of thymic tuft cells in *Kn2* BALB/cByJ × *Aire*^{DTR} C57BL/6 F$_1$ mice disrupts the medullary NKT2–IL-4 axis. a–d, Analysis of thymus from *Kn2* BALB/cByJ × *Aire*^{DTR} C57BL/6 F$_1$ mice or *Aire*^{DTR} transgene-negative F$_1$ controls treated for 9 days with diphtheria toxin. **a**, Immunofluorescence staining of thin thymic sections for KRT10 (red) and DCLK1 (green). $n = 3$ mice; two independent experiments. **b**, Counts of CD11c$^-$CD45$^-$EPCAM$^+$ mTECs and DCLK1$^+$ tuft cells in non-transgenic control thymus ($n = 9$ mice) or F$_1$ thymus ($n = 12$ mice). **c**, Flow cytometry plots gated on TCRβ^{int}CD1d$^+$ iNKT cells showing intracellular PLZF and RORγt staining for iNKT subset analysis. Right, counts of NKT1 (PLZF$^-$RORγt$^-$), NKT2 (PLZF$^+$RORγt$^-$) and NKT17 (PLZF$^-$RORγt$^+$) in non-transgenic control thymus ($n = 35$ mice) or F$_1$ thymus ($n = 23$ mice). **d**, Flow cytometry plots gated on TCRβ$^+$CD8$^+$ single-positive thymocytes showing intracellular EOMES staining. Right, counts of EOMES$^+$ cells in non-transgenic control thymus ($n = 20$ mice) or F$_1$ thymus ($n = 18$ mice). **b–d**, Data are mean ± s.d. Unpaired, parametric, two-tailed Student's *t*-test. Pooled from three independent experiments.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | Characterization of *Pou2f3*$^{-/-}$ mice. **a**, Representative on-edge, full-face mid-thymic sections (5 μm) stained with haematoxylin and eosin showing grossly normal thymic architecture in *Pou2f3*$^{-/-}$ mice. $n = 3$ mice; two independent experiments. **b**, Flow cytometry plots (left) and frequencies (right) of mTEC subsets from C57BL/6 control and *Pou2f3*$^{-/-}$ mice. $n = 5$ mice per genotype; three independent experiments. **c**, Flow cytometry plots (left) and counts (right) of thymocyte subsets from C57BL/6 control and *Pou2f3*$^{-/-}$ mice. $n = 5$ mice per genotype; three independent experiments. **d**, Flow cytometry plots (left) of TCRβ^{int}CD1d^{+} thymic iNKT cells in C57BL/6 control ($n = 26$ mice), *Pou2f3*$^{-/-}$ ($n = 19$ mice) or *Trpm5*$^{-/-}$ ($n = 10$ mice) mice. **e**, Gating strategy of iNKT subset analysis (left) and counts (right) of NKT1 (PLZF^{-}RORγt^{-}) and NKT17 (PLZF^{-}RORγt^{+}) in C57BL/6 control ($n = 15$ mice), *Pou2f3*$^{-/-}$ ($n = 19$ mice) or *Trpm5*$^{-/-}$ ($n = 10$ mice) mice. **f**, Flow cytometry plots gated on TCRβ^{+}CD8^{+} single-positive thymocytes showing intracellular EOMES staining. Quantified in Fig. 4d. **g**, Flow cytometry plots gated on splenic TCRβ^{int}CD1d^{+} iNKT cells showing intracellular PLZF and RORγt staining for iNKT subset analysis, NKT1 (PLZF^{-}RORγt^{-}), NKT2 (PLZF^{+}RORγt^{-}) and NKT17 (PLZF^{-}RORγt^{+}). Right, counts of NKT1 and NKT17 in C57BL/6 control ($n = 5$ mice) or *Pou2f3*$^{-/-}$ ($n = 4$ mice) mice. Two independent experiments. **b**–**e**, **g**, Data are mean ± s.d. **b**, **c**, **g**, Unpaired, parametric, two-tailed Student's *t*-test. **d**, **e**, One-way, non-parametic ANOVA (Kruskal–Wallis test). Pooled from three independent experiments.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0316-7

Histidine catabolism is a major determinant of methotrexate sensitivity

Naama Kanarek[1,2,3,4], Heather R. Keys[1], Jason R. Cantor[1,2,3,4], Caroline A. Lewis[1], Sze Ham Chan[1], Tenzin Kunchok[1], Monther Abu-Remaileh[1,2,3,4], Elizaveta Freinkman[1], Lawrence D. Schweitzer[4] & David M. Sabatini[1,2,3,4]*

The chemotherapeutic drug methotrexate inhibits the enzyme dihydrofolate reductase[1], which generates tetrahydrofolate, an essential cofactor in nucleotide synthesis[2]. Depletion of tetrahydrofolate causes cell death by suppressing DNA and RNA production[3]. Although methotrexate is widely used as an anticancer agent and is the subject of over a thousand ongoing clinical trials[4], its high toxicity often leads to the premature termination of its use, which reduces its potential efficacy[5]. To identify genes that modulate the response of cancer cells to methotrexate, we performed a CRISPR–Cas9-based screen[6,7]. This screen yielded *FTCD*, which encodes an enzyme—formimidoyltransferase cyclodeaminase—that is required for the catabolism of the amino acid histidine[8], a process that has not previously been linked to methotrexate sensitivity. In cultured cancer cells, depletion of several genes in the histidine degradation pathway markedly decreased sensitivity to methotrexate. Mechanistically, histidine catabolism drains the cellular pool of tetrahydrofolate, which is particularly detrimental to methotrexate-treated cells. Moreover, expression of the rate-limiting enzyme in histidine catabolism is associated with methotrexate sensitivity in cancer cell lines and with survival rate in patients. In vivo dietary supplementation of histidine increased flux through the histidine degradation pathway and enhanced the sensitivity of leukaemia xenografts to methotrexate. The histidine degradation pathway markedly influences the sensitivity of cancer cells to methotrexate and may be exploited to improve methotrexate efficacy through a simple dietary intervention.

To identify genes that contribute to the response of cancer cells to methotrexate, we performed a genome-wide, positive-selection CRISPR–Cas9-based screen[6,7] in the erythroleukaemia cell line HEL. We selected this cell line because of its high sensitivity to methotrexate in a competitive growth assay of 42 haematopoietic cell lines in the presence of methotrexate (Fig. 1a, Extended Data Fig. 1a–c).

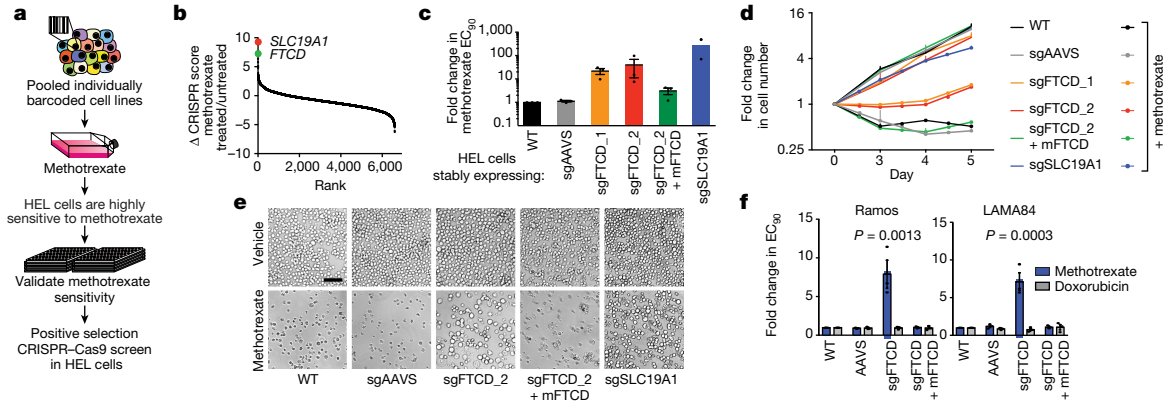


Fig. 1 | Loss of *FTCD* decreases the sensitivity of cancer cells to methotrexate. a, Selection of the HEL cell line for the CRISPR–Cas9-based screen. The genomes of 42 haematopoietic cancer cell lines were individually barcoded. The cell lines were pooled together and treated with 0, 0.1, 0.5 and 5 μM methotrexate for 6 days. Genomic barcodes were sequenced to determine the relative representation of each line in the mixed culture at the various methotrexate concentrations. The erythroleukaemia HEL cell line was identified as a sensitive cell line suitable for a genome-wide, positive-selection CRISPR–Cas9-based screen. **b**, The two top hits in the CRISPR–Cas9-based screen[6,7] were *SLC19A1* and *FTCD*. Genes were ranked by the difference between their CRISPR score[10] in the methotrexate-treated and the vehicle-treated samples. **c**–**e**, Targeting *FTCD* by CRISPR–Cas9 in HEL cells decreased their sensitivity to methotrexate. **c**, Fold change in the EC_{90} of methotrexate in HEL cells that were treated with methotrexate for 5 days and stably expressing the indicated constructs. EC_{90} values for methotrexate are relative to wild-type (WT) cells ($n = 3$, except for SLC19A1 for which $n = 2$, biological replicates). **d**, HEL cells stably expressing the indicated constructs were counted daily to assess their survival after treatment with 5 μM methotrexate ($n = 3$, biological replicates). **e**, Differential interference contrast micrographs of HEL cells stably expressing the indicated constructs and treated with 5 μM methotrexate for 3 days. Scale bar, 100 μm. A representative experiment is presented ($n = 3$, biological replicates). **f**, CRISPR–Cas9-depletion of *FTCD* decreased the sensitivity of additional cell lines (Ramos and LAMA84) to methotrexate. Shown are fold changes in the EC_{90} values of methotrexate and the control drug, doxorubicin, compared to wild-type cells ($n = 3$, biological replicates, ordinary one-way ANOVA, comparing sgFTCD to each of the other samples. For doxorubicin all P values were not significant). In **c**, **d**, **f**, error bars indicate s.e.m. Abbreviations: mFTCD, cells stably expressing mouse *Ftcd* cDNA that is not targeted by the sgRNA used in this experiment; sgAAVS, cells stably expressing an sgRNA targeting the non-coding AAVS locus[6,7]; sgFTCD, cells stably expressing an sgRNA targeting *FTCD*; sgSLC19A1, cells stably expressing an sgRNA targeting *SLC19A1*.

[1]Whitehead Institute for Biomedical Research and Massachusetts Institute of Technology, Department of Biology, Cambridge, MA, USA. [2]Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Koch Institute for Integrative Cancer Research and Massachusetts Institute of Technology, Department of Biology, Cambridge, MA, USA. [4]Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA, USA. *e-mail: sabatini@wi.mit.edu

© 2018 Macmillan Publishers Limited, part of Springer Nature. All rights reserved.

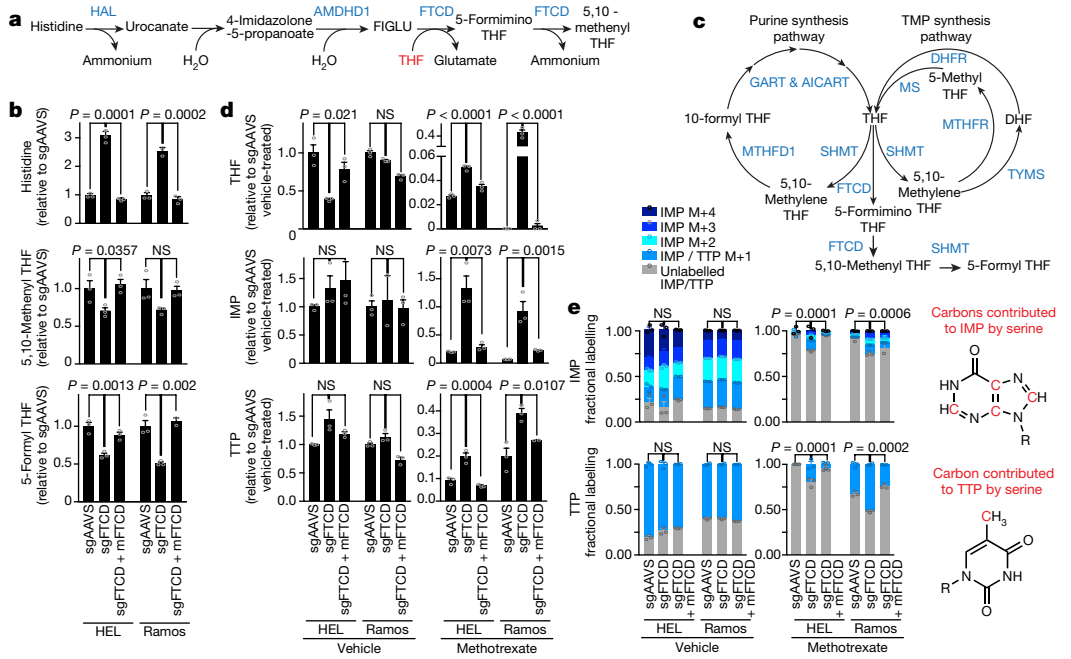


Fig. 2 | FTCD depletion enables cancer cells to maintain THF pools and nucleotide synthesis even when treated with methotrexate. a, The histidine degradation pathway. FTCD has two functions in the histidine degradation pathway. The formimidoyltransferase domain metabolizes THF and the histidine breakdown product FIGLU to produce glutamate and 5-formimino THF, and the cyclodeaminase domain metabolizes 5-formimino THF to 5,10-methenyl THF[14]. **b**, CRISPR–Cas9-induced *FTCD* depletion increased levels of histidine (top), decreased levels of 5-formyl THF (bottom) in HEL and Ramos cells, and decreased levels of 5,10-methenyl THF in HEL cells and slightly decreased levels of 5,10-methenyl THF in Ramos cells (middle). **c**, The use of THF by the purine synthesis pathway (left cycle), the TMP synthesis pathway (right external cycle), the methyl cycle (right internal cycle), and by FTCD. 5,10-Methenyl THF is metabolized to 5-formyl THF by the enzyme serine hydroxymethyltransferase[26–28]. **d**, The greater cellular pool of THF and significantly higher abundance of nucleotides in methotrexate-treated cells after *FTCD* depletion. THF (top), IMP (middle) and TTP (bottom) levels were measured by LC–MS in vehicle- and methotrexate-treated HEL and Ramos cells. **e**, Newly synthesized IMP (top) and TTP (bottom) in *FTCD*-depleted, methotrexate-treated cells shown by [U-^{13}C]serine labelling. IMP and TTP fractional labelling is shown for vehicle-treated and methotrexate-treated HEL and Ramos cells. Right, the four carbons that are contributed to IMP or the one carbon contributed to TTP by [U-^{13}C]serine. For **b**, **d**, **e**, $n = 3$, biological replicates. Statistical tests used were a one-way ANOVA (**b**, **d**) and a two-way ANOVA for the unlabelled fractions (**e**). ANOVA was done as multiple comparisons between sgFTCD and each of the other samples. Error bars indicate s.e.m. Abbreviations: AICART, aminoimidazolecarboxamide ribonucleotide formyltransferase (AICAR formyltransferase); DHFR, dihydrofolate reductase; GART, phosphoribosylglycinamide formyltransferase (GAR transformylase); IMP, inosine monophosphate; MS, methionine synthase; MTHFR, methylene tetrahydrofolate reductase; NS, not significant; SHMT, serine hydroxymethyltransferase; TTP, thymidine triphosphate; TYMS, thymidylate synthase.

We focused our efforts on haematopoietic cell lines because methotrexate is most frequently used to treat haematopoietic malignancies[9]. The two highest-scoring[10] genes in the screen were *SLC19A1* and *FTCD* (Fig. 1b, Extended Data Fig. 1d, e). SLC19A1 is a reduced folate transporter that imports methotrexate into cells, and its depletion reduces methotrexate sensitivity in cultured cells[11] and in patient tumours[12,13].

FTCD is an enzyme involved in histidine catabolism and has not previously been associated with methotrexate sensitivity. Depletion of *FTCD* using two distinct single-guide RNAs (sgRNAs) increased the EC_{90} (90% of maximal effective concentration) of methotrexate by more than tenfold relative to controls (Fig. 1c–e). Importantly, expression of the murine *Ftcd* cDNA, which is resistant to sgRNA-mediated targeting, re-sensitized *FTCD*-depleted cells to methotrexate, validating the results of our screen (Fig. 1c–f). RNA interference (RNAi)-mediated knockdown of *FTCD* also increased the EC_{90} of methotrexate (Extended Data Fig. 1f, g), which suggests that even subtle reductions in *FTCD* expression are sufficient to alter methotrexate sensitivity. CRISPR–Cas9-mediated depletion of *FTCD* also decreased the sensitivity of Ramos (Burkitt's lymphoma) and LAMA84 (chronic myeloid leukaemia) cells to methotrexate (Fig. 1f, Extended Data Fig. 1h), which demonstrates that our findings are generalizable to cell lines derived from additional haematopoietic malignancies.

FTCD catalyses two reactions in the histidine degradation pathway[14] (Fig. 2a, Extended Data Fig. 2a). The formimidoyltransferase domain metabolizes tetrahydrofolate (THF) and the histidine-breakdown product formiminoglutamic acid (FIGLU) to produce glutamate and 5-formimino THF. The cyclodeaminase domain further metabolizes 5-formimino THF to 5,10-methenyl THF[14]. We profiled FTCD-relevant metabolites in *FTCD*-depleted HEL and Ramos cells using liquid chromatography coupled with mass spectrometry (LC–MS). Depletion of *FTCD* increased the levels of histidine (Fig. 2b, top, Extended Data Fig. 2b), and decreased the levels of 5,10-methenyl THF (Fig. 2b, middle, Extended Data Fig. 2b) and the downstream metabolite 5-formyl THF (Fig. 2b, bottom, Extended Data Fig. 2b). The modest decline in 5,10-methenyl THF is probably due to an additional pool of this metabolite, synthesized by the enzyme methylenetetrahydrofolate dehydrogenase 1 (MTHFD1). This pool of 5,10-methenyl THF is directly channelled to 10-formyl THF by MTHFD1 and is unlikely to serve as a substrate for the build-up of 5-formyl THF[15]. These data indicate that depletion of *FTCD* decreases flux through the histidine degradation pathway.

Methotrexate treatment causes a decrease in the cellular pool of THF, which inhibits nucleotide synthesis and is considered to be the main cause of cell death by the treatment[3]. The use of THF by FTCD results in the synthesis of 5-formyl THF, a folate entity that has no described role as an enzyme cofactor and is considered to be a storage form of folate[16] (Fig. 2c, Extended Data Fig. 2d). We proposed that the consumption of THF by FTCD for the synthesis of 5-formyl THF would be particularly harmful to cells in which the THF pool is already limiting, such as in methotrexate-treated cells. Indeed, depletion of the *FTCD* gene resulted in significantly higher levels of THF in methotrexate-treated cells (Fig. 2d, top, Extended Data Figs. 2c, 3c, d), indicating that the use of THF by FTCD contributes to the exhaustion of THF pools observed in wild-type cells treated with methotrexate.

© 2018 Springer Nature Limited. All rights reserved.

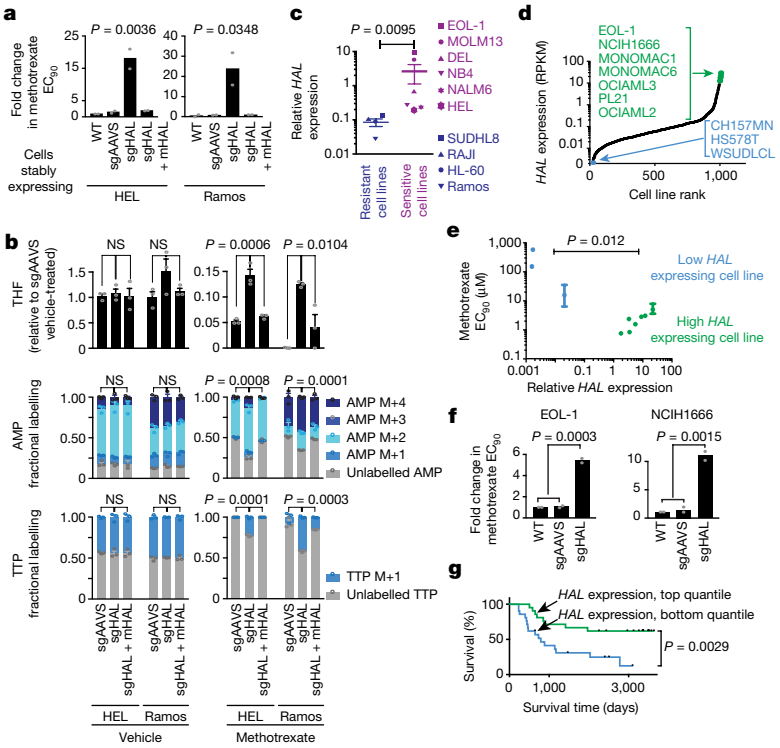


Fig. 3 | The histidine degradation pathway affects the sensitivity of cancer cells to methotrexate and *HAL* expression is associated with treatment response in patients with ALL. a, CRISPR–Cas9-induced *HAL* depletion decreased the sensitivity of HEL and Ramos cells to methotrexate. The average fold change in the EC_{90} of methotrexate relative to the wild-type is shown ($n = 2$, biological replicates, one-way ANOVA, as multiple comparisons between sgHAL and each of the other samples). **b**, The greater cellular pool of THF in methotrexate-treated cells after *HAL* depletion. Top, THF levels were measured by LC–MS in vehicle- and methotrexate-treated HEL and Ramos cells. Newly synthesized AMP (middle) and TTP (bottom) in *HAL*-depleted and control cells, treated with methotrexate, shown by [U-^{13}C]serine labelling ($n = 3$, biological replicates; for THF, one-way ANOVA; for AMP and TTP, two-way ANOVA for the unlabelled fraction). **c**, *HAL* expression is significantly higher in cells that are more sensitive to methotrexate. The response to methotrexate was determined in a pooled fashion using genomically barcoded cell lines (Fig. 1a). *HAL* expression was measured by qPCR ($n = 4$ for resistant cell lines and $n = 6$ for sensitive cell lines, biologically independent samples, two-tailed Kolmogorov–Smirnov test). **d**, **e**, Differences in *HAL* expression are associated with methotrexate sensitivity in cancer cell lines. **d**, Cell lines were ranked by *HAL* expression according to RNA-seq data[21]. Green, cell lines with high *HAL* expression. Blue, cell lines with low *HAL* expression. **e**, *HAL* expression (by qPCR) and methotrexate EC_{90} are shown for each of the coloured cell lines from **d** (low *HAL* ($n = 3$), high *HAL* ($n = 7$), two-tailed Kolmogorov–Smirnov test, individual EC_{90} values: $n = 2$, biological replicates). **f**, CRISPR–Cas9-depletion of *HAL* in EOL-1 and NCIH1666 cells decreased their sensitivity to methotrexate. Fold changes in methotrexate EC_{90} compared to wild type are shown ($n = 2$, biological replicates, one-way ANOVA). **g**, *HAL* expression predicts better survival in paediatric patients with ALL treated with a regimen that included methotrexate. Kaplan–Meier curves of overall survival of patients with ALL with high (top quantile, green) or low (bottom quantile, blue) expression of *HAL*[24] ($n = 21$). Error bars indicate s.e.m.

We observed significantly higher levels of inosine monophosphate (IMP) and thymidine triphosphate (TTP) after methotrexate treatment in *FTCD*-depleted cells compared to control cells (Fig. 2d middle and bottom), which suggests that the increased THF pool size in these cells contributes to nucleotide biosynthesis. To examine the ability of methotrexate-treated, *FTCD*-depleted cells to synthesize nucleotides, we used uniformly labelled (U-) [^{13}C]serine to label newly synthesized nucleotides in cells treated with methotrexate. Loss of FTCD significantly increased the labelling of IMP and TTP in methotrexate-treated cells (Fig. 2e), without changing the overall labelling efficiency (Extended Data Fig. 2e, f) or the cellular pools of glycine and serine (Extended Data Fig. 3a, b). These results suggest that *FTCD* depletion enables the survival of methotrexate-treated cells by sparing cellular THF pools, and thus enabling de novo nucleotide synthesis to continue.

To determine the contribution of other components of the histidine degradation pathway (Fig. 2a) to methotrexate sensitivity, we genetically targeted histidine ammonia lyase (HAL) and amidohydrolase domain containing 1 (AMDHD1), two enzymes of the histidine degradation pathway that function upstream of FTCD[17–19]. Similar to our findings with FTCD, CRISPR–Cas9-mediated depletion of these enzymes decreased the sensitivity of cells to methotrexate (Fig. 3a, Extended Data Fig. 4a). As in *FTCD*-depleted cells, methotrexate-treated, *HAL*-depleted cells had significantly higher levels of THF (Fig. 3b, top), and higher rates of nucleotide synthesis as monitored by [U-^{13}C]serine tracing in comparison to control cells (Fig. 3b, middle and bottom). Notably, whereas control cell lines show increased total serine and [U-^{13}C]serine levels upon methotrexate treatment, *HAL*-depleted HEL cells show total and [U-^{13}C]serine levels that are similar to vehicle-treated cells (Extended Data Fig. 4d, f). The increased serine levels in methotrexate-treated control cells are probably the result of reduced serine usage owing to limiting cellular levels of THF. Glycine labelling appears to be the same in all samples (Extended Data Fig. 4c, e). These data suggest that, as with the depletion of *FTCD*, the depletion of *HAL* results in a higher availability of reduced folate and thus greater nucleotide synthesis rates in cells treated with methotrexate.

We found that the expression of *HAL*—which encodes the rate-limiting enzyme in the histidine degradation pathway[20]—but not of *FTCD* or *AMDHD1* was significantly higher (Fig. 3c, Extended Data Fig. 4b) in haematopoietic cell lines that we identified as being highly sensitive to methotrexate (Fig. 1a, Extended Data Fig. 1a–c). To follow up on this finding, we analysed RNA-seq data from 1,010 cell lines[21] to identify those with the highest and lowest expression of *HAL* (Fig. 3d), and found that cell lines with high *HAL* expression were more sensitive to methotrexate than those with low expression (Fig. 3e). Depletion of *HAL* in two high *HAL*-expressing lines (EOL-1 and NCIH1666) substantially decreased their methotrexate sensitivity (Fig. 3f, Extended Data Fig. 5a), revealing a functional role for HAL in the response of these cells to the drug. Thus, cells with endogenously low or

© 2018 Springer Nature Limited. All rights reserved.

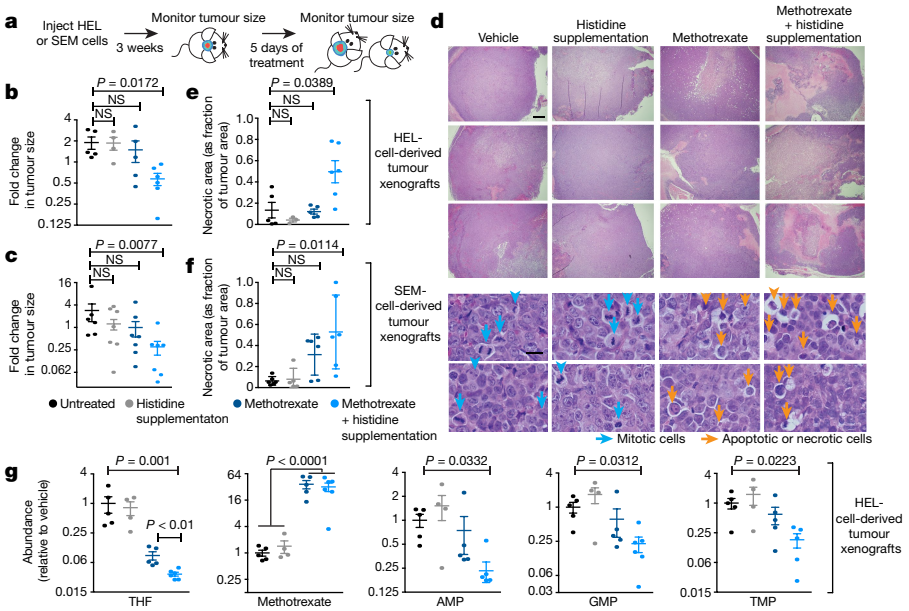


Fig. 4 | In vivo histidine supplementation increases flux through the histidine degradation pathway and sensitizes tumours to methotrexate. **a**, Workflow for the in vivo assessment of the methotrexate-sensitivity of tumour xenografts after histidine supplementation. HEL or SEM cells were injected subcutaneously into NOD-SCID mice. Three weeks later tumour size was assessed by in vivo imaging and mice were randomly divided into four experimental groups: vehicle-treated, histidine supplementation, methotrexate-treated and histidine supplementation combined with methotrexate treatment. The treatments were followed by a second imaging session. **b**, **c**, The combination of methotrexate and histidine supplementation is the only treatment that resulted in a significant reduction in tumour size. Fold changes in tumour size as measured in vivo before and after the different treatments are presented for HEL-cell-derived tumours (**b**), and SEM-cell-derived tumours (**c**) (non-parametric one-way ANOVA, Kruskal–Wallis test, Dunn's post hoc test). **d**, The combination of methotrexate treatment and histidine supplementation caused cancer cell death. Haematoxylin and eosin staining of tumours from mice injected with HEL cells and treated with the different regimes is presented at two magnifications. Mitotic cells (blue arrows) and apoptotic or necrotic cells (orange arrows) are marked in the bottom panel. Scale bars: 200 μm (top), 10 μm (bottom). **e**, **f**, The combination of methotrexate and histidine supplementation was the only treatment that resulted in significant necrosis in tumours. Necrotic areas as detected by haematoxylin and eosin staining (**d**, top three rows, right column) were measured in the different groups (non-parametric one-way ANOVA, Kruskal–Wallis test, Dunn's post hoc test). **g**, THF levels decreased after methotrexate treatment and decreased even further when methotrexate was combined with histidine supplementation. Methotrexate levels were not different in tumours from mice treated with methotrexate alone or in combination with histidine supplementation. Nucleotide abundance was significantly lower in tumours from mice treated with methotrexate and histidine supplementation compared to tumours from vehicle-treated mice (non-parametric one-way ANOVA, Kruskal–Wallis test, Dunn's post hoc test. Methotrexate, two tailed *t*-test). **b**, **d**, **e**, vehicle ($n = 5$), histidine ($n = 4$), methotrexate ($n = 6$), methotrexate + histidine ($n = 6$). **c**, **f**, ($n = 6$), ($n = 7$), ($n = 7$), ($n = 7$) for the same groups. **g**, ($n = 5$), ($n = 4$), ($n = 5$), ($n = 6$), for the same groups. Centre bars indicate mean, error bars indicate s.e.m.

exogenously lowered *HAL* expression are both less sensitive to methotrexate than control cells.

Methotrexate is used as the standard of care in the treatment of the most common paediatric malignancy, acute lymphoblastic leukaemia (ALL)[22], with relatively high success rates that unfortunately are accompanied by severe toxicity[5]. At present, treatment-outcome prediction markers are lacking[23], leading physicians to prescribe methotrexate regimens based on clinical parameters instead of personalized pharmacogenetic factors. Notably, we found that young patients with ALL that showed high *HAL* expression in the ALL cells have significantly higher survival rates, compared to patients in the same study[24] with low *HAL* expression (Fig. 3g). Expression levels of *SLC19A1*, which encodes the transporter for methotrexate, as well as the genes (*AMDHD1* and *FTCD*) encoding histidine degradation pathway enzymes showed no significant association with patient survival (Extended Data Fig. 5b–d). These data warrant further investigation of the expression level of *HAL* as a prediction factor for treatment success rate in patients with ALL that are treated with methotrexate.

The association between expression levels of the rate-limiting enzyme in the histidine degradation pathway and methotrexate sensitivity in cancer cell lines and patients raised the possibility that increasing flux through this pathway could enhance the cytotoxic effects of methotrexate. We proposed that in vivo histidine supplementation would increase histidine degradation and consumption of THF by FTCD, which might enhance the toxic effects of methotrexate on tumour cells. To assess the response of cancer cells to methotrexate in the presence of above-normal levels of histidine, we generated tumours by injecting HEL or SEM cells subcutaneously into NOD-SCID mice and treated the tumour-bearing mice with vehicle, histidine, methotrexate or a combination of histidine and methotrexate (Fig. 4a). To reliably assess the contribution of histidine supplementation, we intentionally used a mild regimen of methotrexate consisting of doses lower than those normally used in mice[25].

As expected, tumours from mice treated with this reduced dosing of methotrexate did not have a significant decrease in size (Fig. 4b, c, Extended Data Fig. 6a, b). However, tumours from mice given both methotrexate and histidine showed a marked decrease in tumour size that was significantly greater than in any of the other treatment cohorts (Fig. 4b, c). Additionally, tumours from mice treated with methotrexate and histidine supplementation appeared histologically worse and contained large necrotic areas and what morphologically appeared to be necrotic or apoptotic cells (Fig. 4d–f, Extended Data Figs. 6c, d, 7a). Organs known to be adversely affected by methotrexate showed no further toxicity in mice treated with the combined therapy compared to mice treated with methotrexate alone (Extended Data Figs. 7b, 9a), and no difference was observed in the weight loss of the mice treated with methotrexate alone compared to methotrexate treatment supplemented with histidine (Extended Data Fig. 9b). No enhanced toxicity was observed in mice treated with the combined therapy compared to methotrexate alone even when the mice were treated for 15 days (Extended Data Figs. 10, 11).

According to our hypothesis, histidine supplementation should increase the flux through the histidine degradation pathway and cause

© 2018 Springer Nature Limited. All rights reserved.

increased consumption of THF, a limiting metabolite in methotrexate-treated tumours. Notably, THF levels were significantly lower in tumours from mice treated with the combined therapy compared to methotrexate alone, which suggests lower THF availability in these tumours, although methotrexate levels were not different (Fig. 4g, Extended Data Fig. 8). Both methotrexate and folate levels were very similar in the plasma of mice treated with methotrexate alone or in combination with histidine supplementation (Extended Data Fig. 9e, f). Metabolites of the histidine degradation pathway were increased in HEL-cell-derived tumours in mice treated with histidine supplementation (Extended Data Fig. 9c, d).

Lower availability of THF should compromise the capacity of cells to synthesize nucleotides, which are essential for cell proliferation and viability. Indeed, the levels of AMP, GMP and TMP in tumours from mice treated with both methotrexate and histidine were significantly lower than those found in tumours from vehicle-treated mice (Fig. 4g, Extended Data Fig. 8). Collectively, our data indicate that histidine supplementation can increase the efficacy of methotrexate in mice in vivo by enhancing the depletion of THF within tumour cells.

Taken together, our results provide evidence for a role of the histidine degradation pathway in the cellular response to the widely used chemotherapy agent methotrexate. Endogenous expression levels of *HAL*, which encodes the rate-limiting enzyme of the pathway, are associated with the sensitivity of cancer cell lines to methotrexate, and the overall survival of patients with ALL who are treated with methotrexate. These findings suggest that *HAL* expression might serve as a clinical predictor for better responders to methotrexate treatment among patients with ALL, and may be informative for decisions regarding therapy strategies. Furthermore, both genetic perturbation and dietary enhancement of the pathway changed the sensitivity of haematopoietic cancer cells to the chemotherapy. As standard protocols for administering methotrexate are often accompanied by severe toxicity[5], we suggest that dietary supplementation with histidine could represent a relatively low-risk intervention that might enable reduced dosing of this toxic agent and therefore a greater clinical benefit.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0316-7.

Received: 27 August 2017; Accepted: 6 June 2018;
Published online 11 July 2018.

1. Farber, S. & Diamond, L. K. Temporary remissions in acute leukemia in children produced by folic acid antagonist, 4-aminopteroyl-glutamic acid. *N. Engl. J. Med.* **238**, 787–793 (1948).
2. Lieberman, I. & Ove, P. Control of growth of mammalian cells in culture with folic acid, thymidine, and purines. *J. Biol. Chem.* **235**, 1119–1123 (1960).
3. Hitchings, G. H. & Burchall, J. J. Inhibition of folate biosynthesis and function as a basis for chemotherapy. *Adv. Enzymol.* **27**, 417–468 (1965).
4. ClinicalTrials.gov https://clinicaltrials.gov/ct2/results?cond=&term=methotrexate&cntry1=&state1=&recrs= (NIH, 2017).
5. Howard, S. C., McCormick, J., Pui, C. H., Buddington, R. K. & Harvey, R. D. Preventing and managing toxicities of high-dose methotrexate. *Oncologist* **21**, 1471–1482 (2016).
6. Shalem, O. et al. Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).
7. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80–84 (2014).
8. Solans, A., Estivill, X. & de la Luna, S. Cloning and characterization of human FTCD on 21q22.3, a candidate gene for glutamate formiminotransferase deficiency. *Cytogenet. Cell Genet.* **88**, 43–49 (2000).
9. Wilson, P. M., Danenberg, P. V., Johnston, P. G., Lenz, H. J. & Ladner, R. D. Standing the test of time: targeting thymidylate biosynthesis in cancer therapy. *Nat. Rev. Clin. Oncol.* **11**, 282–298 (2014).
10. Wang, T. et al. Identification and characterization of essential genes in the human genome. *Science* **350**, 1096–1101 (2015).
11. Assaraf, Y. G. & Schimke, R. T. Identification of methotrexate transport deficiency in mammalian cells using fluoresceinated methotrexate and flow cytometry. *Proc. Natl Acad. Sci. USA* **84**, 7154–7158 (1987).
12. Matherly, L. H. & Taub, J. W. Methotrexate pharmacology and resistance in childhood acute lymphoblastic leukemia. *Leuk. Lymphoma* **21**, 359–368 (1996).
13. Guo, W. et al. Mechanisms of methotrexate resistance in osteosarcoma. *Clin. Cancer Res.* **5**, 621–627 (1999).
14. Mao, Y. et al. Structure of the bifunctional and Golgi-associated formiminotransferase cyclodeaminase octamer. *EMBO J.* **23**, 2963–2971 (2004).
15. Hatefi, Y., Osborn, M. J., Kay, L. D. & Huennekens, F. M. Hydroxymethyl tetrahydrofolic dehydrogenase. *J. Biol. Chem.* **227**, 637–647 (1957).
16. Field, M. S., Szebenyi, D. M. & Stover, P. J. Regulation of de novo purine biosynthesis by methenyltetrahydrofolate synthetase in neuroblastoma. *J. Biol. Chem.* **281**, 4215–4221 (2006).
17. La Du, B. N., Howell, R. R., Jacoby, G. A., Seegmiller, J. E. & Zannoni, V. G. The enzymatic defect in histidinemia. *Biochem. Biophys. Res. Commun.* **7**, 398–402 (1962).
18. HumanCyc https://humancyc.org (2017).
19. Yu, C. et al. The development of PIPA: an integrated and automated pipeline for genome-wide protein function annotation. *BMC Bioinformatics* **9**, 52 (2008).
20. Brand, L. M. & Harper, A. E. Studies on the production and assessment of experimental histidinemia in the rat. *Biochim. Biophys. Acta* **444**, 294–306 (1976).
21. Cancer Cell Line Encyclopedia https://broadinstitute.org/ccle
22. Cooper, S. L. & Brown, P. A. Treatment of pediatric acute lymphoblastic leukemia. *Pediatr. Clin. North Am.* **62**, 61–73 (2015).
23. Kodidela, S., Suresh Chandra, P. & Dubashi, B. Pharmacogenetics of methotrexate in acute lymphoblastic leukaemia: why still at the bench level? *Eur. J. Clin. Pharmacol.* **70**, 253–260 (2014).
24. Roberts, K. G. et al. Targetable kinase-activating lesions in Ph-like acute lymphoblastic leukemia. *N. Engl. J. Med.* **371**, 1005–1015 (2014).
25. Liem, N. L. et al. Characterization of childhood acute lymphoblastic leukemia xenograft models for the preclinical evaluation of new therapies. *Blood* **103**, 3905–3914 (2004).
26. Holmes, W. B. & Appling, D. R. Cloning and characterization of methenyltetrahydrofolate synthetase from *Saccharomyces cerevisiae*. *J. Biol. Chem.* **277**, 20205–20213 (2002).
27. Stover, P. & Schirch, V. Serine hydroxymethyltransferase catalyzes the hydrolysis of 5,10-methenyltetrahydrofolate to 5-formyltetrahydrofolate. *J. Biol. Chem.* **265**, 14227–14233 (1990).
28. Stover, P., Kruschwitz, H. & Schirch, V. Evidence that 5-formyltetrahydropteroylglutamate has a metabolic role in one-carbon metabolism. *Adv. Exp. Med. Biol.* **338**, 679–685 (1993).

Acknowledgements We thank all members of the Sabatini laboratory for their advice and help; W. Chen and L. Shinefeld for helpful comments on the manuscript; M. Pacold for advice; M. Cohen for the *R*-luciferase expression vector; and M. Lazzara, D. Lauffenburger, M. Zaitseva, A. Al-Katib and R. Jensen for providing cell lines. We thank J. Selhub for advice in the folate field. This work was supported by grants from National Institutes of Health/National Cancer Institute (R01 CA129105) to D.M.S. and the Department of Defense (W81XWH-15-1-0337) to E.F. Fellowship support was provided by the European Molecular Biology Organization (EMBO) (Long-Term Fellowship ALTF 350-2012) and the American Association for Cancer Research (16-40-38-KANA) to N.K., by the American Cancer Society (PF-12-099-01-TBG) and the Koch Institute (Ludwig Postdoctoral Fellowship) to J.R.C., and by the EMBO (Long-Term Fellowship ALTF 1-2014) to M.A.-R. Additional support to N.K. was provided by the Women In Science/Revson Foundation Award (Weizmann Institute) and The Advancement of Women in Science Award (The Hebrew University). D.M.S. is an investigator of the Howard Hughes Medical Institute and an American Cancer Society Research Professor.

Reviewer information *Nature* thanks C. Frezza and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions N.K. and D.M.S. formulated the research plan and interpreted experimental results. N.K. designed and performed all experiments with assistance from J.R.C. and M.A.-R. H.R.K. performed data analysis for the CRISPR–Cas9-based screen, Cancer Cell Line Encyclopedia (CCLE) gene expression data and data from patients with ALL. J.R.C. carried out the genomic barcoding of 42 haematopoietic cell lines. E.F. and C.A.L. ran the metabolite-profiling core facility and E.F., C.A.L., S.H.C. and T.K. advised and helped to plan experiments for metabolite detection by LC–MS. S.H.C. and T.K. assisted with LC–MS sample preparation and running of the method. L.D.S. designed and prepared some figures for the manuscript, N.K. and D.M.S. wrote the manuscript, and all authors edited it.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0316-7.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0316-7.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to D.M.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

METHODS

Barcoding and co-culture of 42 haematopoietic cell lines. Cell lines used in this manuscript were authenticated by short tandem repeat analysis and were tested and found negative for mycoplasma. The sources of cell lines were as follows. SKM1, P12-Ichikawa, MOLT4, PF382, SEMK2: T. Look (Dana–Farber Cancer Institute (DFCI)); SUDHL4, HBL1, LY7, U2932: M. A. Shipp (DFCI); EOL-1, CEM, REH: S. A. Armstrong (DFCI); KU812, NOMO1, K562, HEL: J. Griffin (DFCI); 697, NALM6: D. M. Weinstock (DFCI); SUDHL8, LY8: A. G. Letai (DFCI); NCIH1666, D. Lauffenburger (MIT); Mono Mac 6: M. Zaitseva (FDA); WSU-DLCL2: A. M. Al-Katib (Wayne State University School of Medicine); CH-157 MN: R. Jensen (University of Utah). All other cell lines were in the Sabatini laboratory's cell bank and were collected or purchased by former lab members.

Construction of unique barcoded plasmids. The oligonucleotides BC-top and BC-bottom were annealed and phosphorylated (T4 PNK, NEB) to generate a linear DNA fragment containing a 5′-AgeI overhang, a 3′-EcoRI overhang, and a degenerate NNNNNNN 7-mer (N is A, T, G, C). The annealed product was cloned into pLKO to generate a collection of pLKO-BC plasmids containing unique 7-mer barcodes that were subsequently determined by Sanger sequencing, as similarly described elsewhere[29].

BC-top:
CCGGTTTTTAGCATCGCCNNNNNNNCTCGCGGCCGCAGGTCCATG
BC-bottom:
AATTCATGGACCTGCGGCCGCGAGNNNNNNNGGCGATGCTAAAAA

Virus production. HEK-293T cells were co-transfected with a pLKO-BC plasmid, the VSV-G envelope plasmid and the ΔVPR lentiviral packaging plasmid using XTremeGene 9 Transfection Reagent (Roche). Culture medium was exchanged 16 h after transfection with the same medium instead supplemented with 20% inactivated fetal bovine serum (IFS). The virus-containing supernatant was collected 48 h after transfection and passed through a 0.45-μm filter to eliminate cells.

Transduction of cell lines. Cells were pelleted, then seeded at a density of 300,000 cells per ml in 6-well plates in 2 ml of RPMI containing 8 μg ml^{-1} polybrene (EMD Millipore), and then transduced with lentivirus by centrifugation at 1,200*g* for 90 min at 37 °C. After a 24-h incubation, cells were pelleted to remove virus and then re-seeded into fresh culture medium containing puromycin, and selected for 72 h.

Competitive survival of barcoded cell lines. Forty-two haematopoietic cancer cell lines that each carried a stably incorporated specific genomic barcode were pooled together and treated for 6 days with methotrexate (0, 0.1, 0.5 and 5 μM). Genomic DNA (gDNA) samples were collected at days 2, 4 and 6 and barcode abundances were determined by high-throughput sequencing (Illumina MiSeq).

Positive-selection CRISPR–Cas9-based screen. Before the screen, the EC_{90} of methotrexate in HEL cells was determined in culture conditions similar to the screen. Detailed protocols for the CRISPR–Cas9-based screen are described elsewhere[30–32]. In short, the sgRNA library[7] was transformed and propagated for virus production as described above, followed by HEL transduction at an efficiency of 60% to avoid more than one sgRNA in each cell. After selection of transduced cells by puromycin, cells were divided into two replicates each of DMSO- and methotrexate-treated groups and were cultured for 14 days. Cells (100×10^6) were collected for genomic DNA purification and sgRNA library deconvolution by HiSeq deep sequencing. Sequencing reads were aligned to the sgRNA library and the abundance of each sgRNA was calculated[10]. The counts from all replicates of the initial cell populations were combined to generate the initial reference dataset. The counts from each sample were normalized for sequencing depth. sgRNAs with fewer than 40 reads in the initial reference dataset were omitted from downstream analyses. The $\log_2$ fold change in abundance of each sgRNA between the final and initial reference populations was calculated and used to define a CRISPR score for each gene[10]. The CRISPR score is the average $\log_2$ fold change in abundance of all sgRNAs targeting a given gene. Genes targeted by fewer than four distinct sgRNAs in the initial reference dataset were omitted from downstream analyses.

Gene targeting using single sgRNAs. To generate HEL, Ramos, LAMA84, EOL-1 and NCIH1666 cells stably expressing the indicated sgRNAs, the following oligonucleotides were cloned into the pLentiCRISPR V1 viral vector.

Oligonucleotides used for sgRNA cloning. sgAAVS: caccgTCCCCTCCACCCCACAGTG (sense), aaacCACTGTGGGGTGGAGGGGAc (antisense). sgFTCD_1: caccgGCA GGTTGAGCGCGATGCGG (sense), aaacCCGCATCGCGCTCAACCTGCc (antisense). sgFTCD_2: caccgGGAGCCATCACACAGACCCC (sense), aaacGGG GTCTGTGTGATGGCTCCc (antisense). sgSLC19A1: caccgTGCGCGGGTCTA CAACGGCG (sense), aaacCGCCGTTGTAGACCCGCGCAc (antisense). sgHAL: caccgGCAGGCAGCTCACTGAGGGA (sense), aaacTCCCTCAGTGA GCTGCCTGCc (antisense). sgAMDHD1: accgATGGAAATTCACCAGGCCGG (sense), aaacCCGGCCTGGTGAATTTCCATc (antisense).

To generate lentivirus, 3×10^6 HEK-293T cells were seeded in a 10-cm plate in DMEM supplemented with 10% IFS. After 24 h, the cells were transfected with the above sgRNA pLenti-encoding plasmids alongside the ΔVPR envelope and CMV VSV-G packaging plasmids using the XTremeGene 9 Transfection Reagent (Roche). Twelve hours after transfection, the medium was aspirated and replaced with 8 ml of fresh medium. Virus-containing supernatants were collected 48 h after the transfection and passed through a 0.45-μm filter to eliminate cells. The transduction of cells was performed as described above.

Gene knockdowns using shRNAs. Lentiviral shRNAs were obtained from The RNAi Consortium[33]. Lentivirus production and viral transduction of HEL cells were performed as described above, followed by a three-day selection with puromycin. Knockdown was verified 48 h after completion of selection by qPCR.

shRNA sequences. GFP, TGCCCGACAACCACTACCTGA. RFP, CGCGTGATGAACTTCGAGGAC. LacZ, CCAACGTGACCTATCCCATTA. FTCD shRNA 1, CCGGGTAGCTTCCCGACTTAT. FTCD shRNA 2, CGAGAAGGAGAACCTCTTCAT. SLC19A1 shRNA 1, CGACGGTGTTCAGAATGTGAA. SLC19A1 shRNA 2, GATTGCATCTTCTCTGTCTAA.

Methotrexate treatment of cell lines in culture. Unless otherwise indicated, 5 μM methotrexate (Sigma) was added to RPMI supplemented with 10% IFS for at least 3 days. Owing to their relatively low sensitivity to the drug, Ramos cells were treated with 20 μM methotrexate.

Cell survival assays. Cells were seeded at a cell density of 50,000 cells per ml in 96-well plates. Three hours later, cells were treated with 7–10 concentrations of methotrexate or doxorubicin in triplicate. Cells were incubated for 5 days, after which an ATP-based, cell viability assay was performed (CellTiter Glo, Promega). Survival curves were calculated by best-fit analysis of the log of the drug concentration (in μM) to fold change of treated cells over vehicle-treated cells. All survival assays included technical triplicates per sample, per experiment.

Cell counting. Cells were counted using a Coulter Counter (Beckman Z2). The percentage viability was calculated by a FACS-based (BD LSR flow cytometer) viability assay using 7-AAD staining (Thermo Fisher Scientific). Each experiment included technical triplicates per sample.

Metabolite profiling by mass spectrometry. *Sample preparation for polar metabolite detection.* For cells in culture, two million cells were collected, washed once with 0.9% NaCl and resuspended in extraction buffer (80% methanol, 20% H_2O plus isotopically labelled internal standards). Samples were then mixed by vortexing for 10 s and the debris was pelleted by a 10-min spin at 18,000*g*. The supernatant was then transferred to a new tube and dried using a liquid nitrogen dryer.

For tumours, the tumours were collected and immediately snap-frozen in liquid nitrogen. Tumours were thawed by homogenization (Kimble cordless handheld homogenizer) in extraction buffer, followed by the procedure described for cell culture samples. All metabolites measured in tumours were normalized to an average of four amino acids (phenylalanine, leucine, valine and tyrosine) as an internal loading control.

For plasma samples, blood was collected from the submandibular vein into anticoagulant tubes (Microvette, Sarstedt), and immediately placed on ice. Samples were spun for 6 min at 1,000*g* and supernatants were collected. From each sample, 10 μl was collected and used for further LC–MS analysis by extraction in 90 μl extraction buffer and sample preparation as described for cultured cells.

Sample preparation for folate detection. The protocol for folate detection was similar to that previously described[34], and adapted with a minor change: the extraction buffer used was 80% methanol, 20% 12.5 mM sodium ascorbate. All subsequent steps were as described[34]. In brief, samples were washed once (0.9% NaCl), then extracted in extraction buffer (80% methanol, 20% 12.5 mM sodium ascorbate, plus aminopterin as internal standard), mixed by vortexing for 10 s and debris was pelleted by a 10-min spin at 18,000*g*. Supernatants were split for polar metabolite and folate detection and dried using a liquid nitrogen dryer manifold. Polar samples were resuspended in water to be run in the mass spectrometer. Folate samples were resuspended in reconstitution buffer (0.5% ascorbic acid, 1% K_2HPO_4 and 0.5% 2-mercaptoethanol) containing rat serum (Sigma, R9759). The rat serum contains the enzymes required to strip the polyglutamate tail from the measured folate. Rat endogenous folate was removed from the rat serum by activated carbon treatment (Sigma, C9157). Polyglutamate tail removal was carried out for 2 h at 37 °C. After the polyglutamate tail removal, the pH was adjusted to 4 using formic acid and samples were loaded on Bond Elute-pH columns (Agilent, 14102062) for a pH-based clean-up. Columns were eluted with 300 μl elution buffer (50% methanol, 0.25% NH_4OAc, 0.5% 2-mercaptoethanol), followed by drying the samples using a liquid nitrogen dryer manifold and resuspension in 50 μl water.

Polar metabolite profiling by LC–MS. Dried polar samples were resuspended in 50 μl water and 2 μl was injected into a ZIC-pHILIC 150 × 2.1 mm (5 μm particle size) column (EMD Millipore). Analysis was conducted on a QExactive benchtop Orbitrap mass spectrometer equipped with an Ion Max source and a HESI II probe, which was coupled to a Dionex UltiMate 3,000 UPLC system (Thermo Fisher Scientific). External mass calibration was performed using the standard calibration mixture every 7 days. Chromatographic separation was achieved using the following conditions: buffer A was 20 mM ammonium carbonate, 0.1% ammonium hydroxide; buffer B was acetonitrile. The column oven and autosampler tray were

© 2018 Springer Nature Limited. All rights reserved.

held at 25 °C and 4 °C, respectively. The chromatographic gradient was run at a flow rate of 0.150 ml min^{-1} as follows: 0–20 min: linear gradient from 80% to 20% B; 20–20.5 min: linear gradient from 20% to 80% B; 20.5–28 min: hold at 80% B. The mass spectrometer was operated in full-scan, polarity switching mode with the spray voltage set to 3.0 kV, the heated capillary held at 275 °C, and the HESI probe held at 350 °C. The sheath gas flow was set to 40 units, the auxiliary gas flow was set to 15 units, and the sweep gas flow was set to 1 unit. The data acquisition was performed over a range of 70–1,000 *m*/*z*, with the resolution set at 70,000, the automatic gain control target at 10e6, and the maximum injection time at 20 ms. Relative quantitation of polar metabolites was performed with XCalibur QuanBrowser 2.2 (Thermo Fisher Scientific) using a 5 p.p.m. mass tolerance and referencing an in-house library of chemical standards.

Histidine was normalized to isotopically labelled histidine as an internal standard; IMP, AMP and TTP levels were normalized to isotopically labelled glutamate as an internal standard.

Detection of folate species by LC–MS. Detection of folate species was performed on the same instrumentation described above. In general, instrument settings remained the same unless specified. After removal of the polyglutamate tails from folate species[34], samples were resuspended in 50 μl water and 15 μl was injected onto an Ascentis Express C18 HPLC column (2.7 μm × 15 cm × 2.1 mm; Sigma Aldrich). The column oven and autosampler tray were held at 30 °C and 4 °C, respectively. The following conditions were used to achieve chromatographic separation: buffer A was 0.1% formic acid; buffer B was acetonitrile with 0.1% formic acid. The chromatographic gradient was run at a flow rate of 0.250 ml min^{-1} as follows: 0–5 min: gradient was held at 5% B; 5–10 min: linear gradient of 5% to 36% B; 10.1–14.0 min: linear gradient from 36–95% B; 14.1–18.0 min: gradient was returned to 5% B. The mass spectrometer was operated in full-scan, positive ionization mode. Data acquisition was performed using three narrow-range scans: 438–450 *m*/*z*; 452–462 *m*/*z*; and 470–478 *m*/*z*, with the resolution set at 70,000, the automatic gain control target at 10e6, and the maximum injection time of 150 ms. Relative quantitation of folate species was performed with XCalibur QuanBrowser 2.2 (Thermo Fisher Scientific) using a 5 p.p.m. mass tolerance. Non-glutamated folate species were identified using chemical standards (see below), together with comparison of experimental MS2 spectra with in silico predicted spectra (Supplementary Fig. 1) Mirror plots were created using R and structures were identified based on the known parental structure (HMDB, www.hmdb.ca), using a combination of ChemDraw (PerkinElmer), CFM:ID[35] (cfmid.wishartlab.com), METLIN (metlin.scripps.edu) and Sirius[36,37] (version 3.5.1).

The standards for different forms of folate are as follows: aminopterin (Sigma, A3411); folate (Schircks Laboratories, 16.203); THF (Schircks Laboratories, 16.207), 5-formyl THF (Schircks Laboratories, 16.221); triglutamate 5-formyl THF (Schircks Laboratories, 16.283); 5,10-methenyl THF (Schircks Laboratories, 16.230); 5-methyl THF (Schircks Laboratories, 16.235). Levels of all folate forms, as well as methotrexate were normalized to aminopterin as an internal standard.

Mass spectrometry reagents. Chloroform was purchased from Sigma-Aldrich and was HPLC grade (ethanol-stabilized). Acetonitrile was purchased from Merck Millipore and was LC–MS Hypergrade. All other solvents were purchased from Fisher and were Optima LC–MS grade. Folate standards were purchased from Schircks laboratories. A mix of 17 isotopically labelled amino acids (Cambridge Isotope Laboratories) was used as the internal standard. FIGLU standard was purchased from Dalton Research Molecules (DC-001551).

Labelling with [U-^{13}C]serine. Cells (500,000) were seeded in RPMI containing 10% dialysed IFS and cultured for 24 h. For FTCD-null cells (Fig. 2e), HEL cells were treated with 5 μM methotrexate for 48 h, and Ramos cells were treated with 20 μM methotrexate for 72 h to decrease THF levels and nucleotide synthesis in wild-type cells. The medium was then replaced with [U-^{13}C]serine-containing medium by counting 1×10^6 cells that were washed once with PBS, and reseeded in RPMI containing 10% dialysed IFS that contained [U-^{13}C]serine and no unlabelled serine. Cells were incubated in [U-^{13}C]serine plus methotrexate for an additional 24 h followed by cell collection and LC–MS analysis. For *HAL*-null cells (Fig. 3b), Ramos cells were treated with 20 μM methotrexate for 48 h before [U-^{13}C]serine labelling to decrease THF levels and nucleotide synthesis in wild-type cells. The medium was then replaced with [U-^{13}C]serine-containing medium by counting 1×10^6 cells that were washed once with PBS, and reseeded in RPMI containing 10% dialysed IFS that contained [U-^{13}C]serine and no unlabelled serine. Cells were incubated in [U-^{13}C]serine plus methotrexate for an additional 24 h followed by cell collection and LC–MS analysis. HEL cells were treated with 5 μM methotrexate for 24 h, in parallel with the [U-^{13}C]serine labelling. Vehicle-treated cells were treated with equivalent concentrations of DMSO for the same times. Natural abundances of labelled nucleotides were corrected as previously described[38,39].

Quantitative PCR. Cell pellets were washed once with PBS before RNA purification using Trizol (ambion). The reverse transcriptase reaction was performed using 2 μg RNA from each sample (SuperScript III, Thermo Fisher Scientific). Quantitative PCR (qPCR) reactions were performed with SYBR green reagent (Roche) and according to the manufacturer's instructions. The QuantStudio 6 Flex (Applied Biosystems) qPCR instrument was used for all reactions. The reference genes used in each qPCR run were *UBC* and *HPRT1*.

Primers used for qPCR. FTCD: forward primer: TGCGAGAAGGAGAACCTCTT, reverse primer: CTCGATGATCCGCTCCTTAG. *HAL:* forward primer: GGCTG AAGGTGGTCTGAACT, reverse primer: AGTCAACAGACGAGGGATGG. *AMDHD1:* forward primer: AGTCTCACACACACGGATCG, reverse primer: GCCT CCGAACTGGTAAATCA. *SLC19A1:* forward primer: GAGAGCTTCA TCACCCCCTA, reverse primer: GGCCAGGTAGGAGTACGACA. Mouse *Ftcd*: forward primer: AACCTGCTAAGCACCAAGGA, reverse primer: GGACC TTTTTCAGACGTCCA. mouse *Hal*: forward primer: GGTGGCCTT AGAGGACAATG, reverse primer: GCTCCCGGTATTTGCTGTAG. Mouse *Amdhd1*: forward primer: TCCACGAGTTTGCAATGAAG, reverse primer: CTCCACCGTGAAGTTGATCC. *GFP:* forward primer: ACAACGTCTA TATCATGGCCGA, Reverse primer: GAAGTTCACCTTGATGCCGTTC. *UBC:* Forward primer: ATTTGGGTCGCGGTTCTT, Reverse primer: TGCCTTGACATTCTCGATGGT. *HPRT1:* Forward primer: TGAC ACTGGCAAAACAATGCA, Reverse primer: GGTCCTTTTCACCAGCAAGCT.

Mouse tumour xenografts. HEL and SEM cells were transduced with *R*-luciferase expressing vector (FUW-RLuc-T2A-PuroR), and selected with puromycin. Cells (2×10^6) were injected subcutaneously into 50 NOD-SCID mice (HEL) or 30 NOD-SCID mice (SEM). Three weeks after injection, tumour size was imaged using the IVIS Spectrum in vivo imaging system (PerkinElmer) after intraperitoneal injection of 50 μg per mouse of the water-soluble substrate Coelenterazine (Nanolight Technologies). A total of 24 mice (HEL) or 27 mice (SEM) with similar tumour sizes were divided into four experimental groups: vehicle-treated (saline), histidine supplementation (400 μl of histidine solution of 46 mg ml^{-1}, histidine maximum solubility at 25 °C), methotrexate-treated (50 mg kg^{-1} methotrexate, based on the weight measured at the day of first imaging), and histidine supplementation combined with methotrexate treatment. Mice were then injected daily for 5 days. All solutions were warmed to 37 °C before injection to reduce stress. After five days of treatment, tumour size was monitored, and tissues were collected for LC–MS and histology. The Committee for Animal Care at the Massachusetts Institute of Technology (MIT) approved all animal procedures carried out in this study. Mice were euthanized when tumours reached the size limit of 1 cm or if other symptoms indicative of low health status appeared. These included lethargy, drop in weight (10% over a day or 20% since the beginning of the experiment), reduced appetite, lack of movement, hunched back and hair loss.

We have complied with all relevant ethical regulations while conducting animal experiments. Sample size was determined as the minimal number of mice absolutely necessary for reliable, significant results (no less than 6 and no more than 8 per cohort, depending on the availability of age-matched mice in the colony). Mice were distributed randomly between the different groups. Injections and imaging were all done by N.K. and were not done blindly. Figure analysis of necrotic areas in haematoxylin and eosin (H&E) slides was done blindly. For all H&E results, different mice within each group showed similar results. Mice assigned to the experiments: NOD-SCID mice - NOD.CB17-Prkdcscid/NCrCrl (originally purchased from Charles River and maintained at the Whitehead mouse facility by self-breeding). HEL-cell-derived tumours experiment: 24 females, ages 8–11.5 weeks, mean 10.7 weeks. SEM-cell-derived tumours experiment: 26 females, ages 12–18 weeks, mean 14 weeks. HEL-cell-derived tumours long-term experiment: mice 1–22 females, mice 23–31 males, ages 6–10 weeks, mean 10 weeks.

Plasmid construction. Cloning into pLentiCRISPR plasmids is described above. Mouse cDNA constructs (mouse *Ftcd*, mouse *Hal* and mouse *Amdhd1*) were cloned into the backbone plasmid pWPI (a contribution from the Didier Trono laboratory, Addgene plasmid number 12254) that was cut at the Pme1 site, and PCR-amplified mouse cDNA was inserted by ligation. The primers used for each gene incorporated a 5′ kozak sequence and Pme1 sites flanking the cloned gene on the 5′ and 3′ sides and are listed below. The luciferase-expression plasmid used for in vivo imaging is FUW-RLuc-T2A-PuroR, which is an FUW lentiviral backbone with the *R*-luciferase gene inserted downstream to a constitutive UBC promoter. The *R*-luciferase gene was cloned using EcoRI. The *R*-luciferase ORF is 936 base-pairs long, missing its stop codon and is in frame with the T2A-PuroR.

Primers used for the cloning. Mouse *Ftcd*: forward primer: GTACGTTTAAACCT CGAGGCCACCATGTCTCAGTTGGTGGAATG, reverse primer: TTCAG TTTAAACTCACTCCTTTCGGGCT. Mouse *Hal*: forward primer: GTACGTT TAAAC CTCGAGGCCACCATGCCTAGGTACACAGTGCA, reverse primer: TTCAGTTTAAAC TTAAAGATCGTCAGACTCTGGG. Mouse *Amdhd1*: forward primer: GTACGTTTAAAC CTCGAGGCCACCATGACGAGCAGCCACC, reverse primer: TTCAGTTTAAAC TCACTTCTTGTAGATGACTTTTCC. Pme site: GTTTAAAC, kozak sequence: CTCGAGGCCACC.

Plasmids used in this study. pLentiCRISPR sgAAVS (Addgene 83906), pLentiCRISPR sgFTCD_1 (Addgene 102312), pLentiCRISPR sgFTCD_2 (Addgene 102313), pLentiCRISPR sgSLC19A1 (Addgene 102314), pLentiCRISPR sgHAL

© 2018 Springer Nature Limited. All rights reserved.

(Addgene 102315), pLentiCRISPR sgAMDHD1 (Addgene 102316), pWP1_mFTCD (Addgene 102317), pWP1_mHAL (Addgene 102318), pWP1_mAMDHD1 (Addgene 102319), FUW-RLuc-T2A-PuroR (Addgene 102320).

Patient survival curves. Patient phenotypic and RNA-seq data for *HAL*, *FTCD*, *AMDHD1* and *SLC19A1* from the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) initiative managed by the NCI were downloaded from the UCSC Xena Genomics Browser (http://xena.ucsc.edu) and the DCC Open sections of the TARGET Data Matrix (https://ocg.cancer.gov/programs/target/data-matrix) on August 15, 2017. Information about TARGET can be found at http://ocg.cancer.gov/programs/target. We then focused on a particular patient subset: the TARGET Substudy Acute Lymphoblastic Leukaemia (ALL) Expansion Phase 2 (phs000218.v18.p7), and further centred our analysis on patients treated with the COG AALL0232 high-risk ALL protocol (ClinicalTrials.gov Identifier NCT00075725). For patients with multiple represented tumour samples, only the first sample was used for downstream analyses. Overall survival was plotted for the upper and lower quartile of *HAL*, *FTCD*, *AMDHD1* and *SLC19A1* expression. Survival curves were compared by the log-rank test and *P* values indicated.

Kaplan–Meier curves were plotted using Prism version 7.0b. End points for patients include death and dropout of patients for reasons other than death, in which cases survival data are integrated in the curve up to the last day of follow-up. The *P* value was calculated using a log-rank Mantel–Cox test.

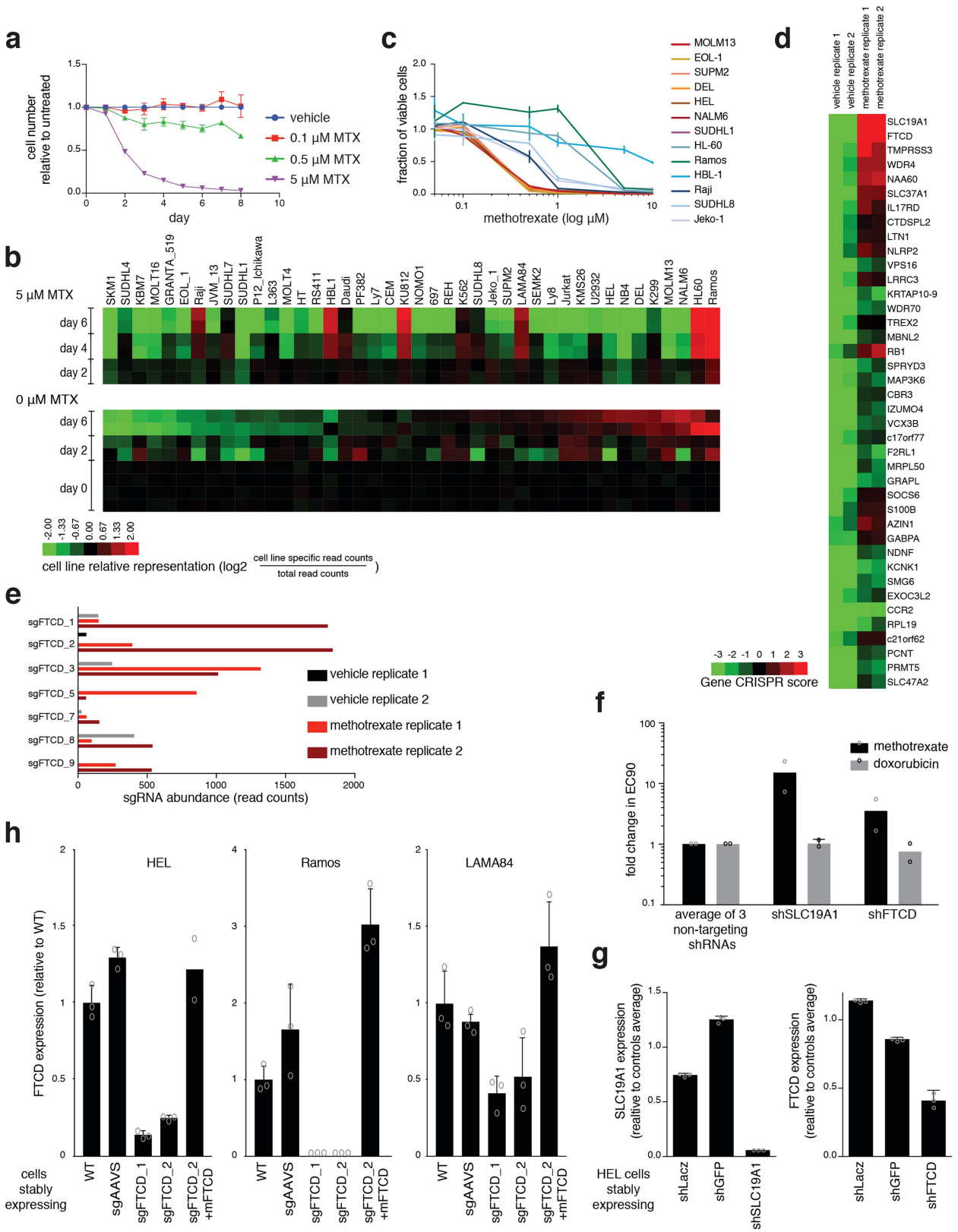
We have complied with all relevant ethical regulations while conducting this patient-derived data collection.

Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. The authors declare that all data generated or analysed during this study are included in this published version and its Supplementary Information Files. Source Data are provided with the online version of the paper. All other datasets generated and/or analysed in the current study are available from the corresponding author on reasonable request.

29. Birsoy, K. et al. Metabolic determinants of cancer cell sensitivity to glucose limitation and biguanides. *Nature* **508**, 108–112 (2014).
30. Wang, T., Lander, E. S. & Sabatini, D. M. Single guide RNA library design and construction. *Cold Spring Harb. Protoc.* **2016**, pdb.prot090803 (2016).
31. Wang, T., Lander, E. S. & Sabatini, D. M. Viral packaging and cell culture for CRISPR-based screens. *Cold Spring Harb. Protoc.* **2016**, pdb.prot090811, (2016).
32. Wang, T., Lander, E. S. & Sabatini, D. M. Large-scale single guide RNA library construction and use for CRISPR–Cas9-based genetic screens. *Cold Spring Harb. Protoc.* **2016**, pdb.top086892 (2016).
33. The RNAi Consortium. The RNAi Consortium shRNA Library https://www.broadinstitute.org/rnai-consortium/rnai-consortium-shrna-library (2017).
34. Chen, L., Ducker, G. S., Lu, W., Teng, X. & Rabinowitz, J. D. An LC-MS chemical derivatization method for the measurement of five different one-carbon states of cellular tetrahydrofolate. *Anal. Bioanal. Chem.* **409**, 5955–5964 (2017).
35. Allen, F., Pon, A., Wilson, M., Greiner, R. & Wishart, D. CFM-ID: a web server for annotation, spectrum prediction and metabolite identification from tandem mass spectra. *Nucleic Acids Res.* **42**, W94–W99 (2014).
36. Böcker, S. & Dührkop, K. Fragmentation trees reloaded. *J. Cheminform.* **8**, 5 (2016).
37. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl Acad. Sci. USA* **112**, 12580–12585 (2015).
38. Fernandez, C. A., Des Rosiers, C., Previs, S. F., David, F. & Brunengraber, H. Correction of ^{13}C mass isotopomer distributions for natural stable isotope abundance. *J. Mass Spectrom.* **31**, 255–262 (1996).
39. Lewis, C. A. et al. Tracing compartmentalized NADPH metabolism in the cytosol and mitochondria of mammalian cells. *Mol. Cell* **55**, 253–263 (2014).

© 2018 Springer Nature Limited. All rights reserved.

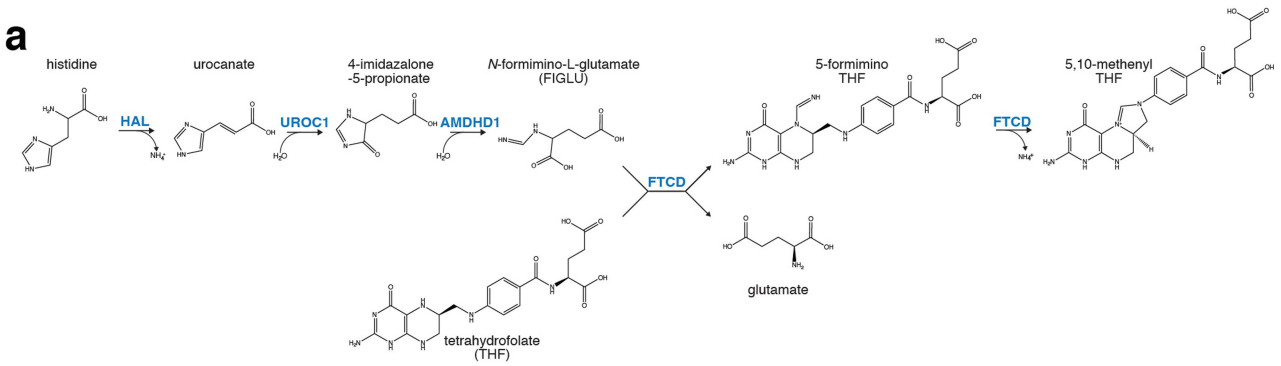


Extended Data Fig. 1 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

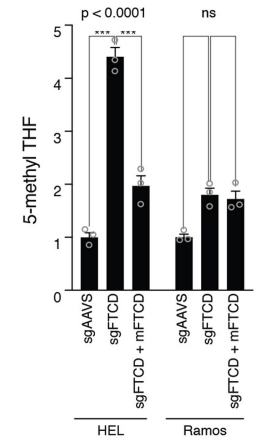
Extended Data Fig. 1 | Loss of *FTCD* decreases the sensitivity of cancer cells to methotrexate. a, Relative daily cell counts of the mixed culture of 42 genomically barcoded haematopoietic cancer cell lines. We focused our efforts on haematopoietic cell lines because methotrexate is most commonly used to treat haematopoietic malignancies[9]. Cells were co-cultured and treated with three concentrations of methotrexate (0.1, 0.5, or 5 μM). Cell counts are presented relative to the vehicle-treated co-culture. $n = 3$, biological replicates. Error bars indicate s.d. **b**, Sequencing results of the competitive co-culture experiment. Heat map of relative barcode abundance for each indicated cell line after treatment with 5 μM methotrexate or vehicle after 2, 4, or 6 days compared to those in the initial cultures. **c**, Validation of the results of the competitive co-culture experiment as shown by individual dose–response curves of seven relatively sensitive cell lines (MOLM13, EOL-1, SUPM2, DEL, HEL, NALM6 and SUDHL1), and six relatively resistant cell lines (HL-60, Ramos, HBL-1, Raji, SUDHL8 and Jeko-1). $n = 2$, biological replicates. Each biological replicate included technical triplicate. Error bars indicate s.d. **d**, Top 40 genes that scored in the genome-wide CRISPR–Cas9-based positive selection screen. These genes had the largest differential CRISPR scores between vehicle- and methotrexate-treated samples. The heat map represents the CRISPR score of two biological replicates of each screen (vehicle- and methotrexate-treated). **e**, Abundance of each of the individual sgRNAs targeting *FTCD* in the screen. Read counts for each sgRNA is presented for each of the screen biological replicates of the vehicle-treated and methotrexate-treated samples. **f**, Fold change in the EC_{90} values of methotrexate and doxorubicin in HEL cells stably expressing shRNA targeting either *SLC19A1* or *FTCD* compared to the average of three non-targeting shRNAs (shGFP, shRFP and shLacZ). $n = 2$, biological replicates. Each biological replicate included technical triplicate. Error bars indicate s.e.m. **g**, Expression levels of *SLC19A1* (left) and *FTCD* (right) in HEL cells stably expressing non-targeting shRNAs (shLacZ and shGFP) or targeting shRNAs (shSLC19A1 and shFTCD). Expression levels were measured by qPCR and normalized to the average of two control genes, *UBC* and *HPRT1*. $n = 3$, technical replicates. Error bars indicate s.e.m. **h**, Validation of genetic depletion of *FTCD* by the CRISPR–Cas9 system in HEL, Ramos and LAMA84 cell lines. Expression of *FTCD* was measured by qPCR and normalized to the average of two control genes, *UBC* and *HPRT1*. $n = 3$ (except for HEL cells expressing sgFTCD_2 + mFTCD, then $n = 2$ owing to loss of RNA from one sample), biological replicates. Error bars indicate s.d.

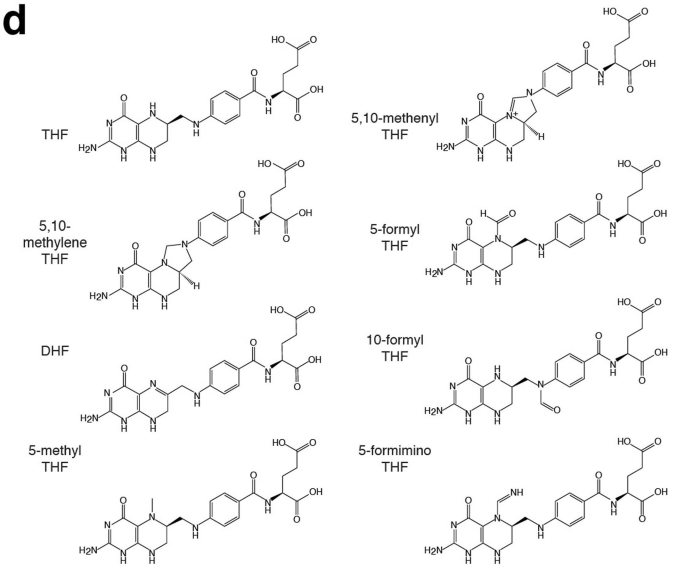
© 2018 Springer Nature Limited. All rights reserved.

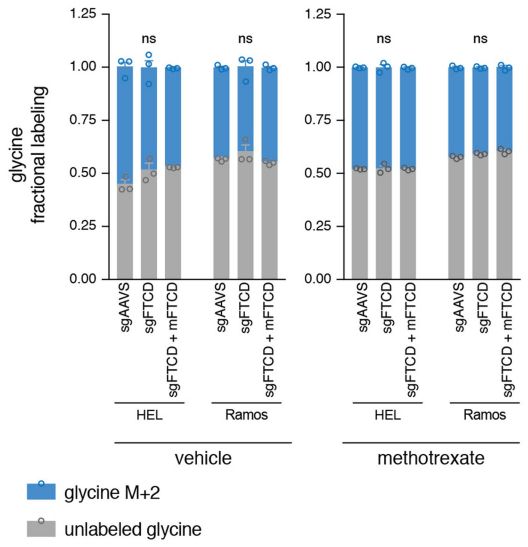


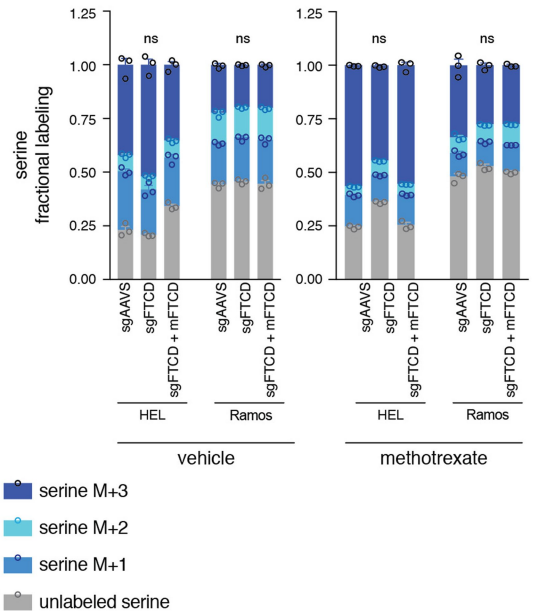
b

metabolite	retention time pHILIC	retention time C18
histidine	14.40	
5,10 methenyl THF		5.40
5 formyl THF		6.56
5 methyl THF		4.47
THF		3.43
IMP	11.05	
TTP	11.54	







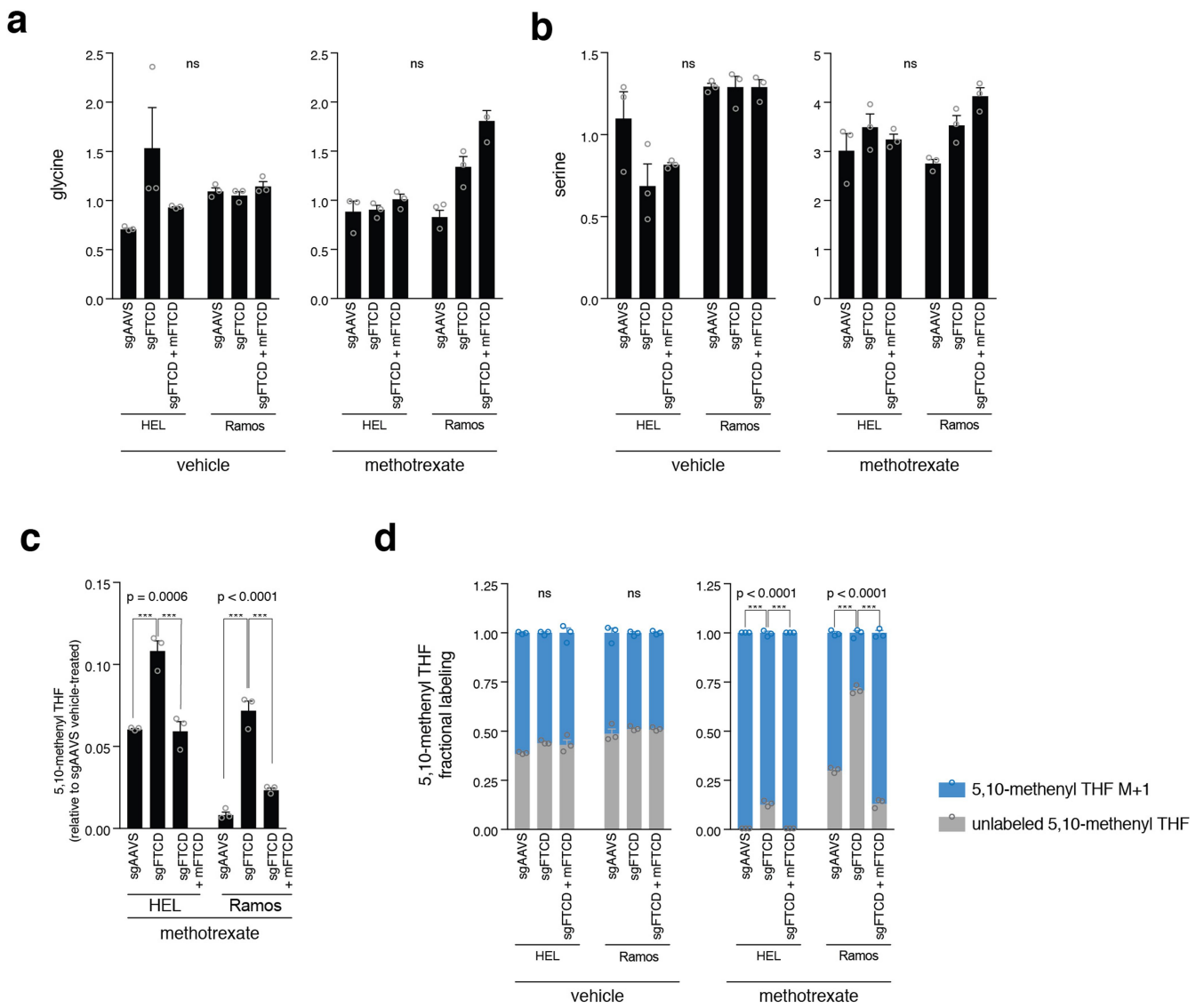


Extended Data Fig. 2 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

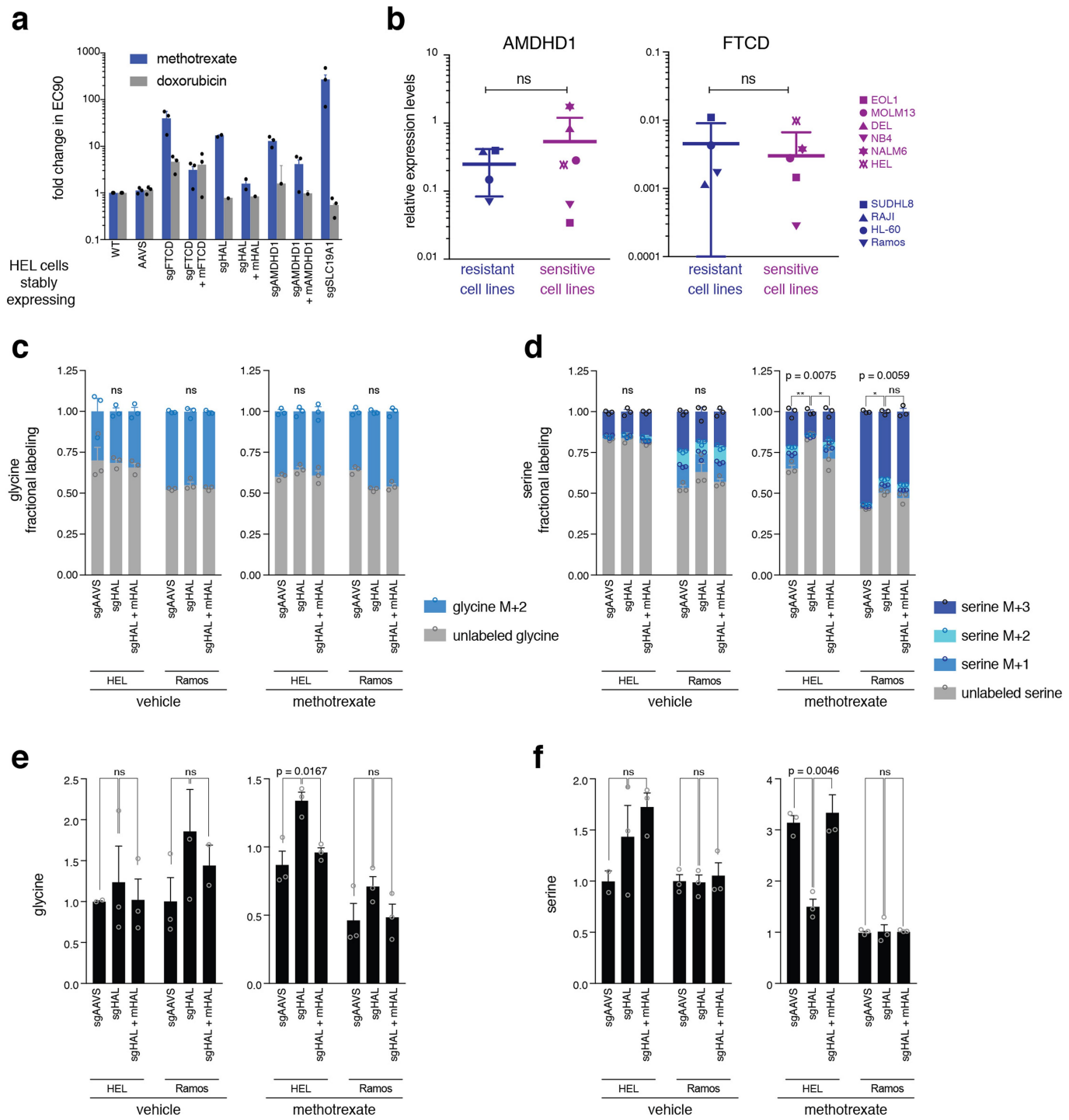
Extended Data Fig. 2 | FTCD depletion enables cancer cells to maintain THF pools and nucleotide synthesis even when treated with methotrexate (part 1). a, The histidine degradation pathway, as previously described[14,17–19]. Enzymes are marked in blue. **b**, Metabolites detected by LC–MS and their corresponding retention times. The LC column used for the detection of each metabolite is indicated. **c**, Greater pool of 5-methyl THF in vehicle-treated HEL cells after *FTCD* depletion. It is not readily clear why, in vehicle-treated HEL cells, *FTCD* depletion caused a reduction in THF levels (Fig. 2d). However, 5-methyl THF levels in these cells were increased significantly (see also Extended Data Fig. 3c), indicating that there was no depletion in the overall recyclable amounts of THF in these cells. 5-Methyl THF levels were measured by LC–MS in vehicle-treated HEL and Ramos cells. 5-Methyl THF levels were normalized to aminopterin as an internal standard. *P* values were calculated using one-way ANOVA. $n = 3$, biological replicates. Error bars indicate s.e.m. **d**, Chemical structures of the folate entities found in cells and mentioned in Fig. 2c. **e**, **f**, The labelling rate by [U-^{13}C]serine is not different between *FTCD*-depleted cells and control cells. Fractional labelling of glycine (**e**) and serine (**f**) is unchanged by *FTCD* depletion in HEL and Ramos cells in vehicle- and methotrexate-treated cells. Glycine and serine levels were normalized to isotopically labelled glutamate as an internal standard. *P* values were calculated for the unlabelled fraction by one-way ANOVA. $n = 3$, biological replicates. Error bars indicate s.e.m. UROC1, urocanate hydratase 1.

© 2018 Springer Nature Limited. All rights reserved.



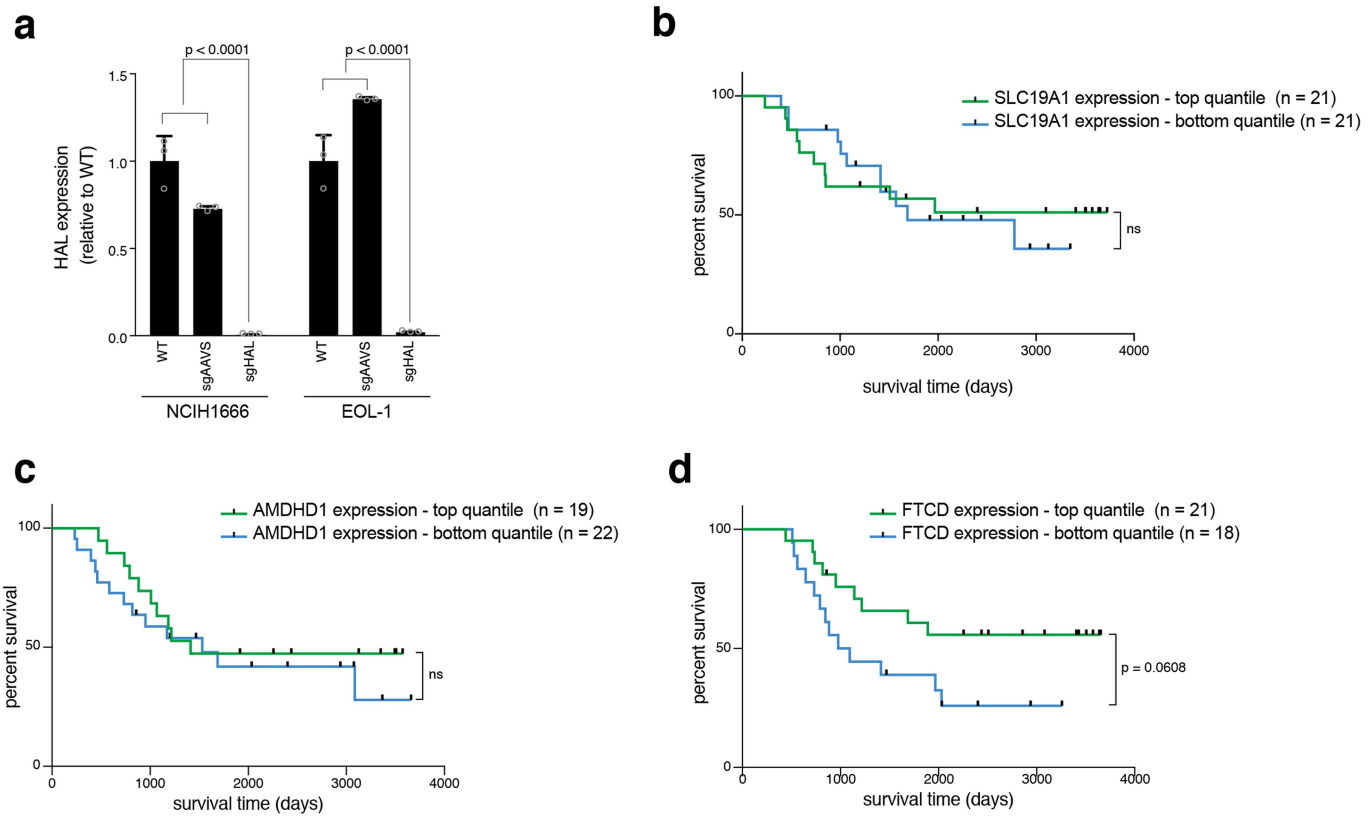
Extended Data Fig. 3 | *FTCD* depletion enables cancer cells to maintain THF pools and nucleotide synthesis even when treated with methotrexate (part 2). a, **b**, Pool sizes of glycine (**a**) and serine (**b**) are not significantly different between *FTCD*-depleted and control cells. Glycine and serine levels were measured in vehicle- or methotrexate-treated HEL and Ramos cells. *P* values were calculated by one-way ANOVA. $n = 3$, biological replicates. Error bars indicate s.e.m. **c**, **d**, Greater cellular pool of 5,10-methenyl THF in methotrexate-treated cells after *FTCD* depletion. [U-^{13}C]serine labelling of 5,10-methenyl THF showed higher abundance of unlabelled 5,10-methenyl THF in *FTCD*-depleted cells, compared to control cells. This suggests higher availability of 5,10-methenyl THF in these cells at the time of labelling onset, which agrees with the higher levels of both THF and 5,10-methenyl THF in the methotrexate-treated *FTCD*-depleted cells compared to the methotrexate-treated control cells. **c**, 5,10-Methenyl THF levels were measured by LC–MS in vehicle- and methotrexate-treated HEL and Ramos cells. 5,10-Methenyl THF levels were normalized to aminopterin as an internal standard. *P* values were calculated by one-way ANOVA. $n = 3$ biological replicates. Error bars indicate s.e.m. **d**, HEL cells were treated with 5 μM methotrexate for 48 h, and Ramos cells were treated with 20 μM methotrexate for 72 h to decrease THF levels and nucleotide synthesis in wild-type cells. The medium was then replaced with [U-^{13}C]serine-containing medium, and cells were incubated in [U-^{13}C]serine plus methotrexate for an additional 24 h followed by cell collection and LC–MS analysis. The higher abundance of unlabelled 5,10-methenyl THF suggests higher availability of reduced folate at the time of labelling onset. 5,10-Methenyl THF levels were normalized to aminopterine as an internal standard. *P* values were calculated for the unlabelled fraction by one-way ANOVA. $n = 3$, biological replicates. Error bars indicate s.e.m.

© 2018 Springer Nature Limited. All rights reserved.



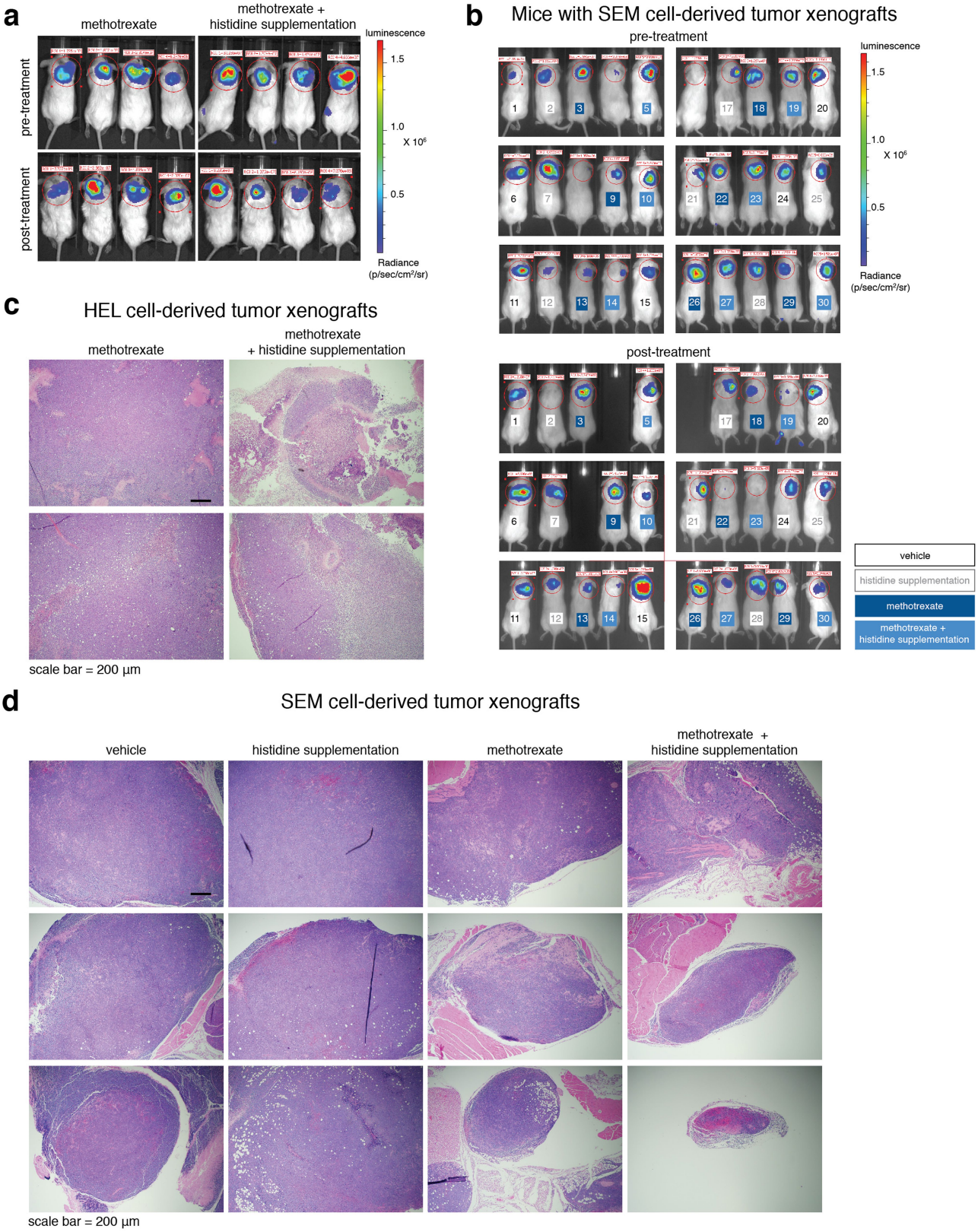
Extended Data Fig. 4 | The histidine degradation pathway affects the sensitivity of cancer cells to methotrexate and *HAL* expression is associated with treatment response in patients with ALL (part 1). **a**, Genetic depletion of the enzymes HAL and AMDHD1 decreased sensitivity to methotrexate, but not to a control drug, doxorubicin. Cell viability after treatment with varying concentrations of methotrexate and doxorubicin was used to calculate the EC_{90} values. $n = 3$, biological replicates. Error bars indicate s.e.m. **b**, Expression levels of *AMDHD1* (left) and *FTCD* (right) were not significantly different across methotrexate-sensitive haematopoietic cell lines compared to methotrexate-resistant cell lines. The response to methotrexate was determined in a pooled fashion using genomically barcoded cell lines (Fig. 1a and Extended Data Fig. 1a–c). Expression levels of *AMDHD1* and *FTCD* were measured by qPCR and normalized to the average of control genes (*UBC* and *HPRT1*). *P* values were calculated using the Kolmogorov–Smirnov test. $n = 4$ for resistant cell lines and $n = 6$ for sensitive cell lines (biologically independent samples). Each qPCR included three technical replicates. Centre bars indicate mean, error bars indicate s.e.m. **c**. Fractional labelling of glycine by [U-^{13}C]serine is unchanged by HAL depletion in HEL and Ramos cells in vehicle-and methotrexate-treated cells. **d**, Uptake of [U-^{13}C]serine is higher in methotrexate-treated control cells but not in methotrexate-treated HAL-deficient cells. **e**, Glycine levels are not significantly different between HAL-deficient and control cells except for HEL cells treated with methotrexate. **f**, Serine levels are not significantly different between HAL-deficient and control cells except for HEL cells treated with methotrexate, in which HAL-deficient HEL cells have similar levels of serine to vehicle-treated cells. For **c**–**f**, glycine and serine levels were normalized to isotopically labelled glutamate as an internal standard. *P* values were calculated for the unlabelled fraction (**c**, **d**) or for total values (**e**, **f**) by one-way ANOVA. $n = 3$, biological replicates. Error bars indicate s.e.m.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 5 | The histidine degradation pathway affects the sensitivity of cancer cells to methotrexate and *HAL* expression is associated with treatment response in patients with ALL (part 2). **a**, Validation of genetic depletion of *HAL* by the CRISPR–Cas9 system in NCIH1666 and EOL-1 cells. Expression levels of *HAL* were measured by qPCR and normalized to the average of two control genes, *UBC* and *HPRT1*. *P* values were calculated by two-way ANOVA. $n = 3$ technical replicates. Error bars indicate s.d. **b**–**d**, Expression levels of *SLC19A1* (**b**), *AMDHD1* (**c**) or *FTCD* (**d**) do not predict survival of paediatric patients with ALL that are treated with a regimen that includes methotrexate. Kaplan–Meier curves of overall survival of patients with ALL with high (top quantile, coloured green) and low (bottom quantile, coloured blue)[24] expression of each of the assayed genes. Patient sample size for each group is indicated. *P* values were calculated using the log-rank (Mantel–Cox) test.

© 2018 Springer Nature Limited. All rights reserved.

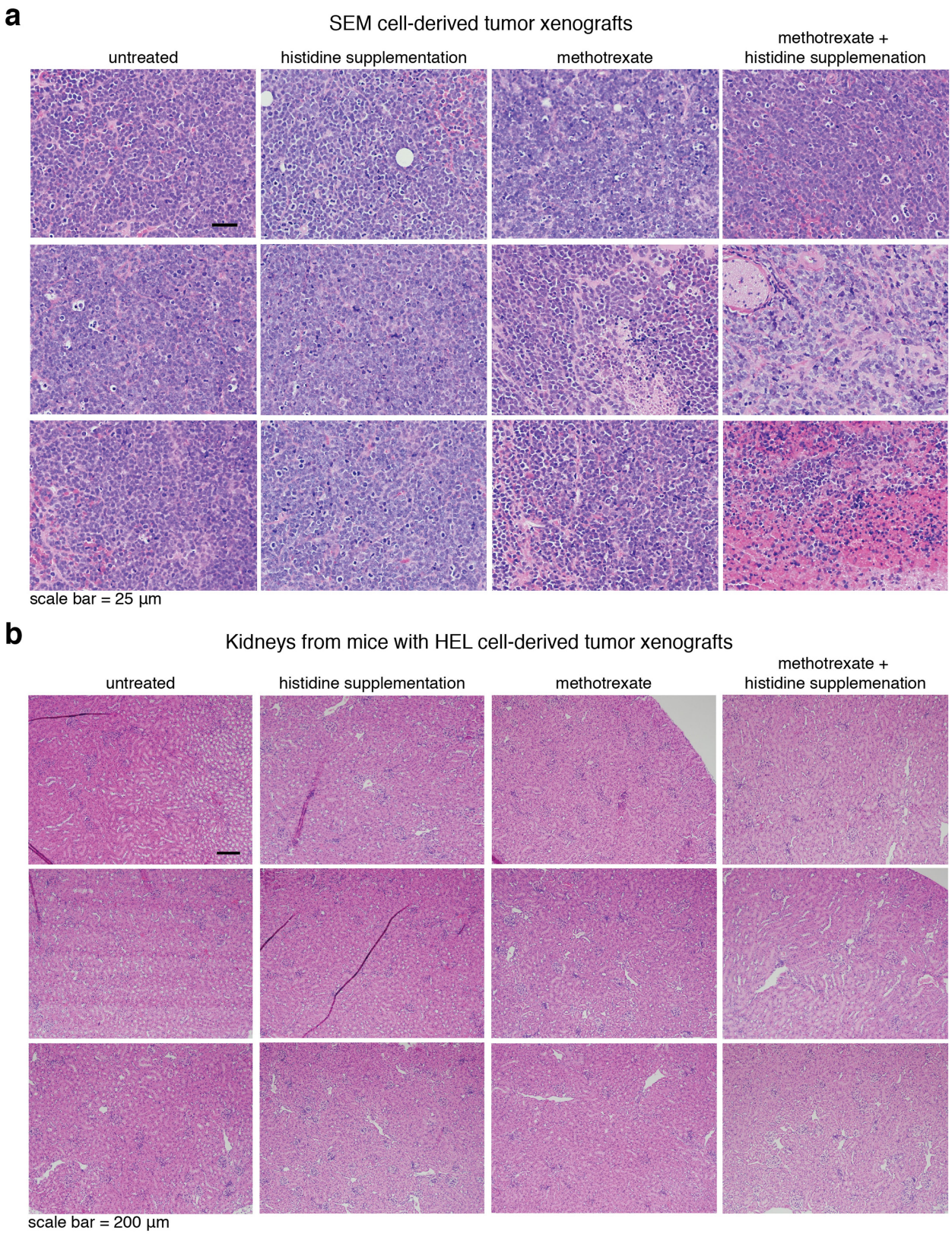


Extended Data Fig. 6 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

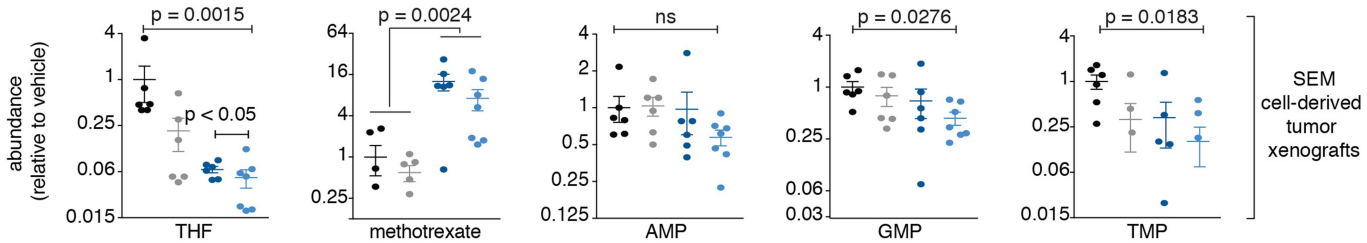
Extended Data Fig. 6 | In vivo histidine supplementation increases flux through the histidine degradation pathway and sensitizes tumours to methotrexate (part 1). a, In vivo imaging of luciferase-expressing HEL tumour xenografts. Mice were imaged before (top images) and after (bottom image) five days of methotrexate treatment alone or in combination with histidine supplementation. For **a**, **c**, HEL-cell-derived tumour-bearing mice: vehicle ($n = 5$), histidine supplementation ($n = 4$), methotrexate ($n = 6$), methotrexate + histidine supplementation ($n = 6$). Four mice per group are presented. **b**, In vivo imaging of luciferase-expressing SEM tumour xenografts. Mice were imaged before (top images) and after (bottom image) five days of treatment. All mice are presented. Mice are numbered in colour by their experimental group. For **b**, **d**, SEM-cell-derived tumour-bearing mice: vehicle ($n = 6$), histidine supplementation ($n = 7$), methotrexate ($n = 7$), methotrexate + histidine supplementation ($n = 7$). **c**. Additional images of H&E analyses of HEL-cell-derived tumour sections from methotrexate-treated and methotrexate + histidine supplementation-treated mice. **d**. H&E images of SEM-cell-derived tumour sections from all treatment groups. Three mice per group are presented.

© 2018 Springer Nature Limited. All rights reserved.



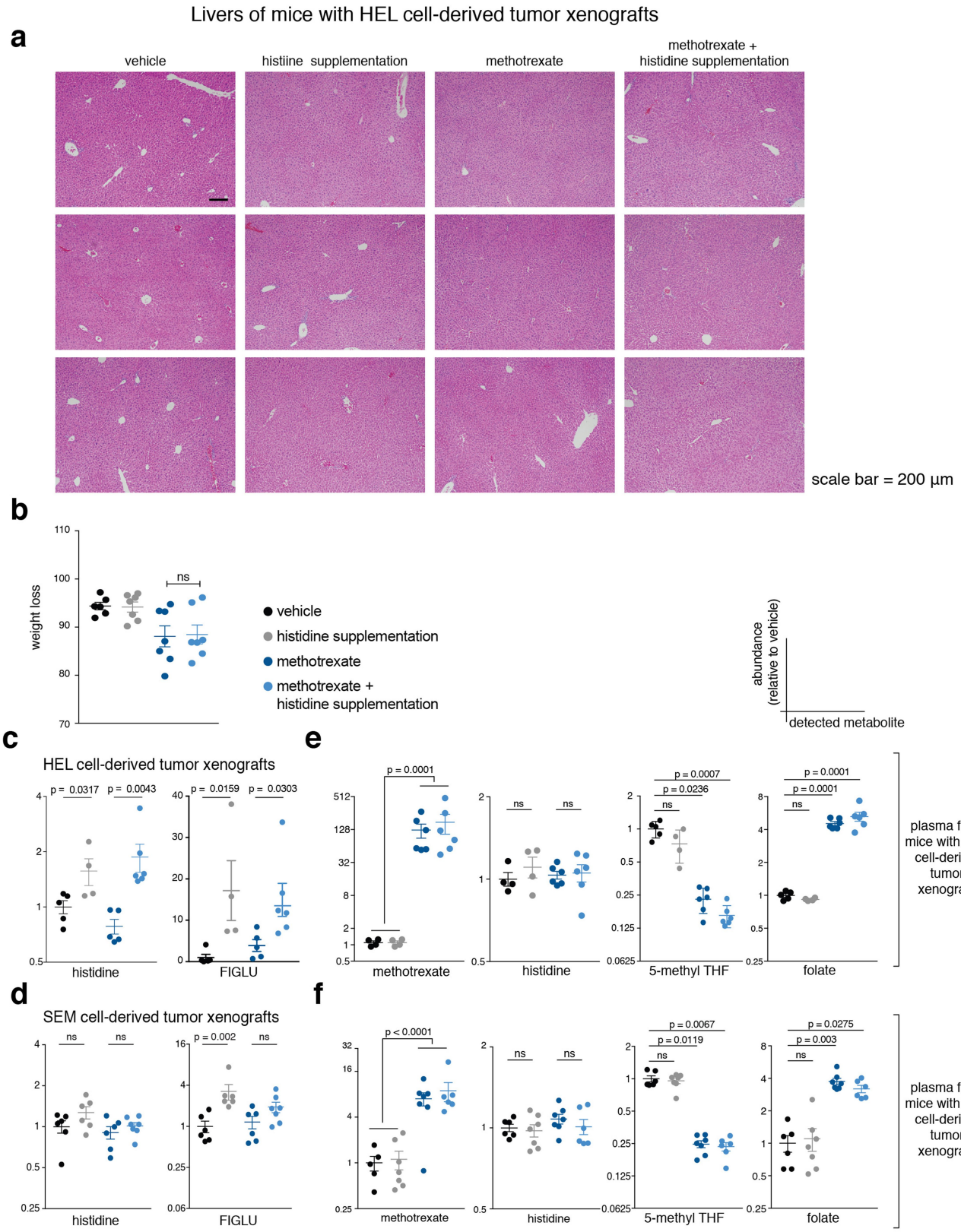
Extended Data Fig. 7 | In vivo histidine supplementation increases flux through the histidine degradation pathway and sensitizes tumours to methotrexate (part 2). a, Higher magnification of H&E images of tumour sections from SEM-cell-derived tumour-bearing mice from all groups. For SEM-cell-derived tumour-bearing mice: vehicle ($n = 6$), histidine supplementation ($n = 7$), methotrexate ($n = 7$), methotrexate + histidine supplementation ($n = 7$). **b**, H&E analyses of kidney sections from HEL-cell-derived tumour-bearing mice from all experimental groups. For HEL-cell-derived tumour-bearing mice: vehicle ($n = 5$), histidine supplementation ($n = 4$), methotrexate ($n = 6$), methotrexate + histidine supplementation ($n = 6$).

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 8 | In vivo histidine supplementation increases flux through the histidine degradation pathway and sensitizes tumours to methotrexate (part 3). THF levels decreased after methotrexate treatment and decreased even further when methotrexate was combined with histidine supplementation. Methotrexate levels were not different in tumours from mice treated with methotrexate alone or in combination with histidine supplementation. Nucleotide abundance was significantly lower in tumours from mice treated with methotrexate and histidine supplementation compared to tumours from vehicle-treated mice. *P* values were calculated using non-parametric one-way ANOVA for all comparisons (Kruskal–Wallis test, Dunn's post hoc test), except for methotrexate, for which *P* values were calculated by a two tailed *t*-test. Vehicle ($n = 6$), histidine ($n = 6$), methotrexate ($n = 6$), methotrexate + histidine ($n = 7$). Centre bars indicate mean, error bars indicate s.e.m.

© 2018 Springer Nature Limited. All rights reserved.

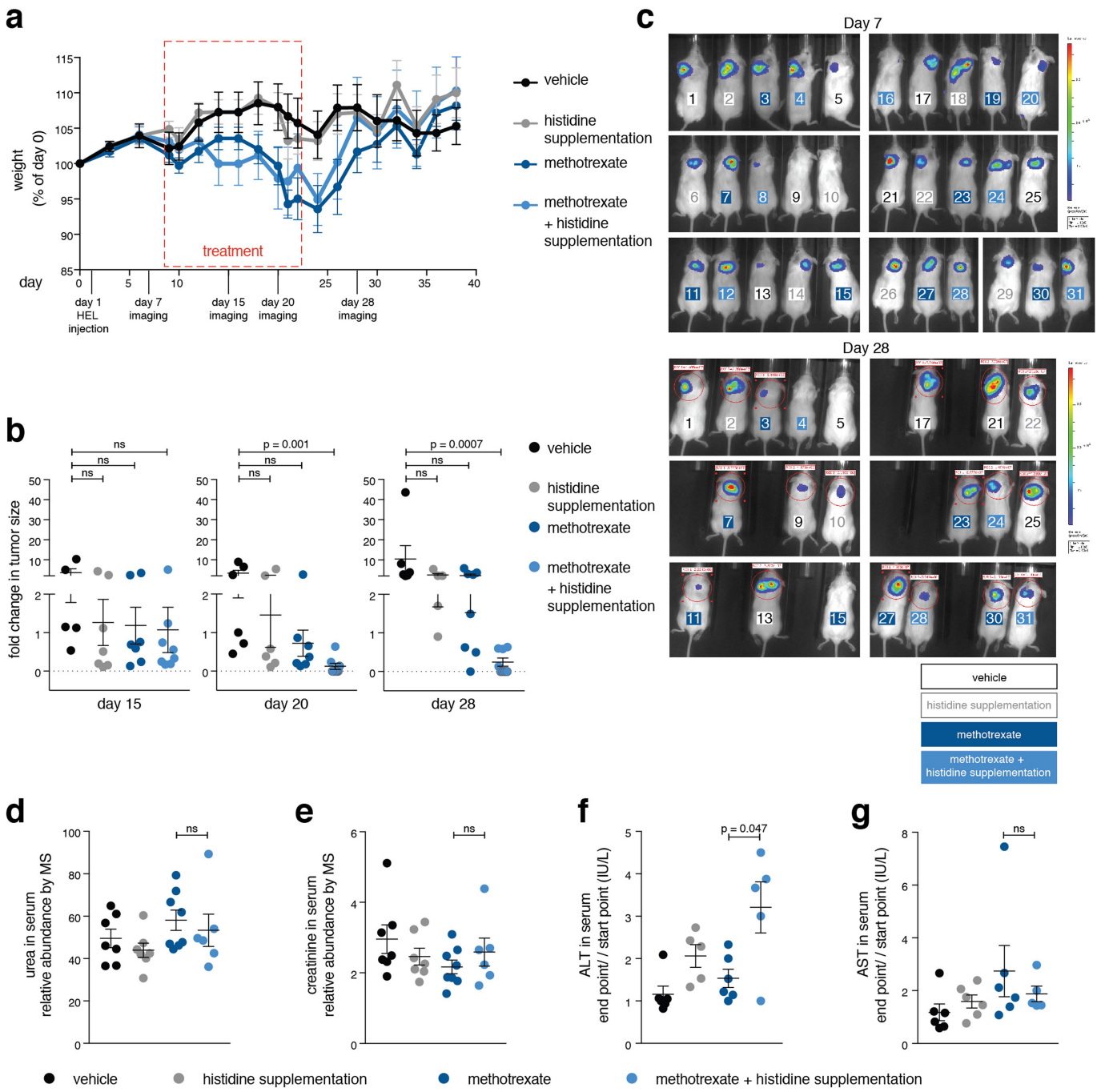


Extended Data Fig. 9 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 9 | In vivo histidine supplementation increases flux through the histidine degradation pathway and sensitizes tumours to methotrexate (part 4). a, H&E analyses of liver sections from HEL-cell-derived tumour-bearing mice from all experimental groups. For HEL-cell-derived tumour-bearing mice: vehicle ($n = 5$), histidine supplementation ($n = 4$), methotrexate ($n = 6$), methotrexate + histidine supplementation ($n = 6$). **b**, Weight loss, presented as a percentage of the weight measured on the first experimental day, before treatment. For SEM-cell-derived tumour-bearing mice: vehicle ($n = 6$), histidine supplementation ($n = 7$), methotrexate ($n = 7$), methotrexate + histidine supplementation ($n = 7$). *P* value was calculated using a non-parametric *t*-test (Mann–Whitney). **c**, **d**, Metabolites of the histidine degradation pathway were increased in HEL tumours from mice treated with histidine supplementation. Histidine (left) and FIGLU (right) levels were measured in tumours by LC–MS and normalized to isotopically labelled histidine as an internal standard. *P* values were calculated using a non-parametric *t*-test (Mann–Whitney). All metabolites measured in tumours were normalized to an average of four amino acids (phenylalanine, leucine, valine and tyrosine) as an internal loading control. For HEL-cell-derived tumour-bearing mice: vehicle ($n = 5$), histidine supplementation ($n = 4$), methotrexate ($n = 5$), methotrexate + histidine supplementation ($n = 6$). For SEM-cell-derived tumour-bearing mice: vehicle ($n = 6$), histidine supplementation ($n = 6$), methotrexate ($n = 6$), methotrexate + histidine supplementation ($n = 7$). **e**, **f**, Plasma levels of methotrexate, histidine, 5-methyl THF and folate. Methotrexate was detected in the plasma of methotrexate-treated mice only (left). No significant difference in histidine levels was detected (second from left). Levels of 5-methyl THF were significantly lower in the plasma of methotrexate-treated mice (second from right). Folate levels increased in the plasma of methotrexate-treated mice (right). Metabolite levels were measured from fresh plasma samples by LC–MS and normalized to isotopically labelled histidine or to aminopterin as an internal standard. *P* values were calculated using non-parametric one-way ANOVA. Group sizes are the same as in **c**, **d**. Centre bars indicate mean, error bars indicate s.e.m.

© 2018 Springer Nature Limited. All rights reserved.

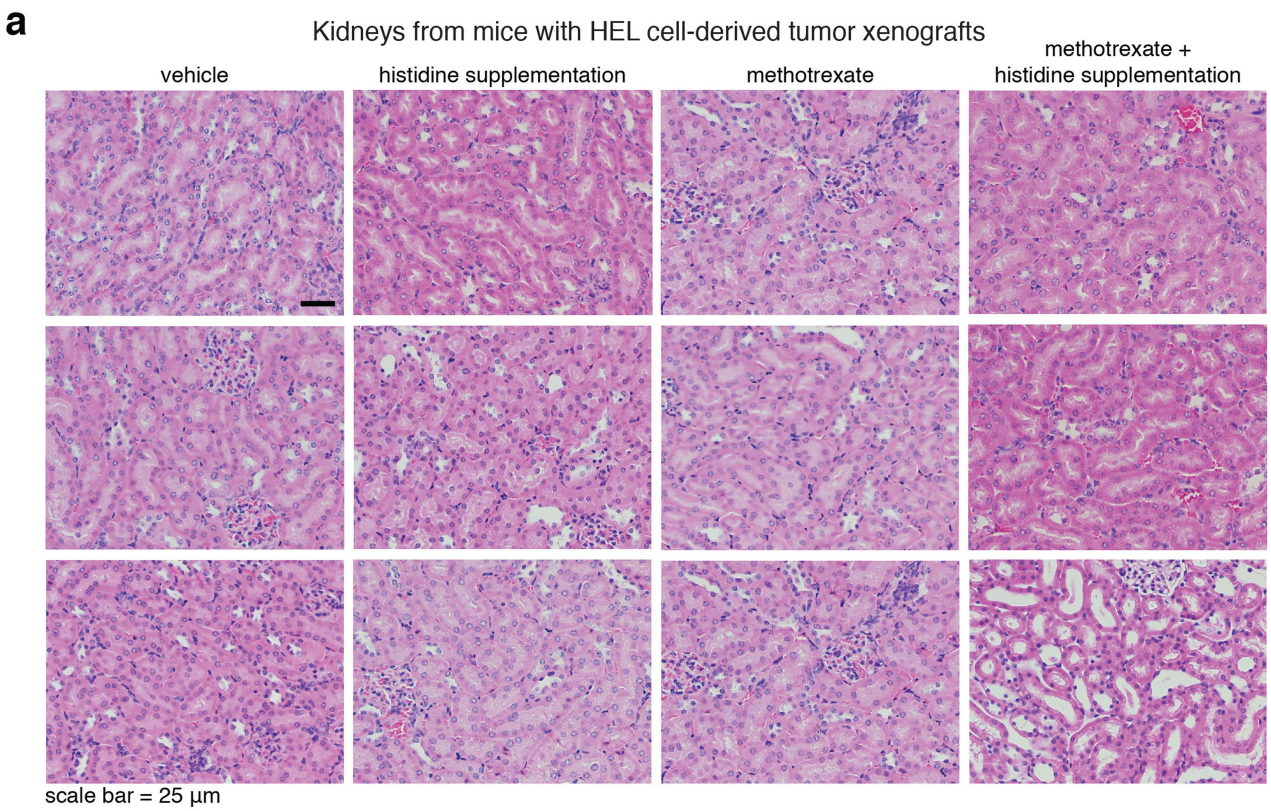


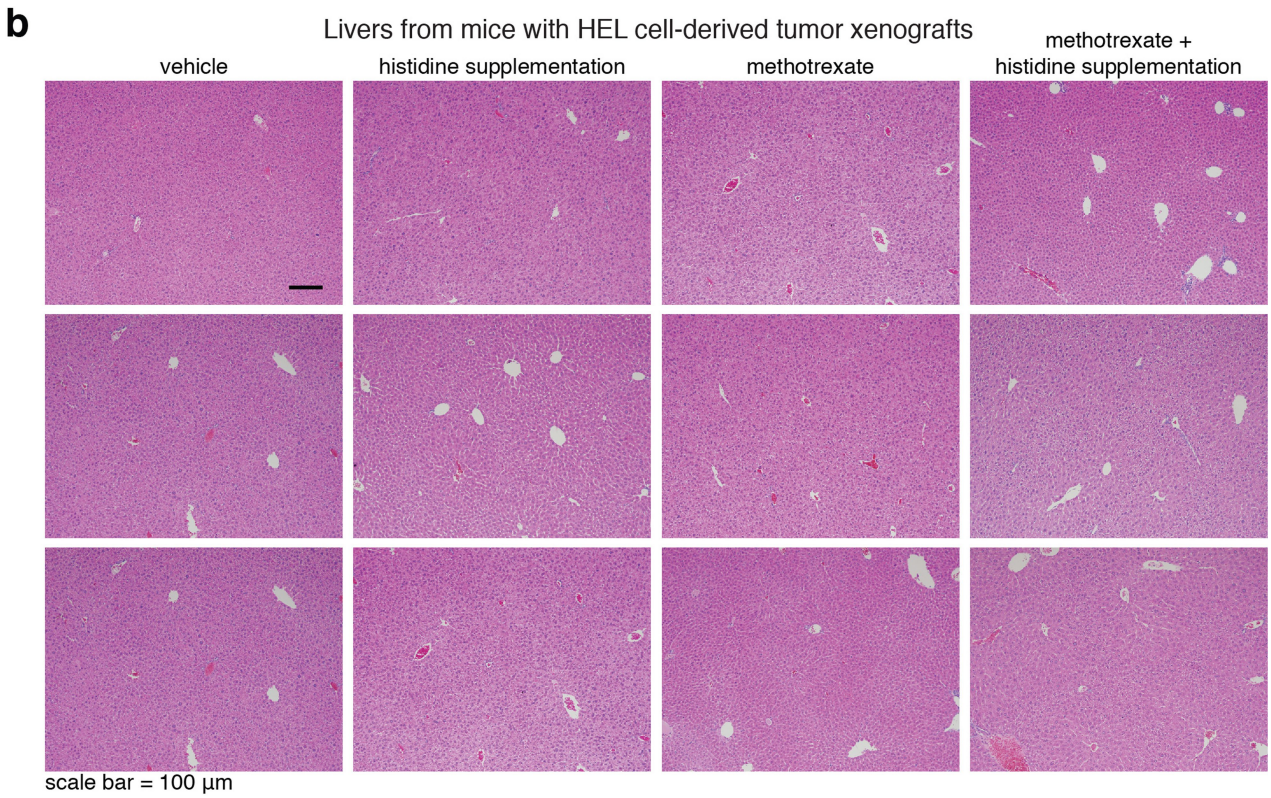
Extended Data Fig. 10 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 10 | In vivo histidine supplementation sensitizes tumours to methotrexate without enhancement of treatment toxicity (part 1). a, We evaluated whether methotrexate treatment combined with histidine supplementation might be more toxic than methotrexate alone by setting up a longer treatment regime of 15 days with a recovery period of two weeks. During the experiment we monitored weight loss and observed no difference between mice treated with methotrexate alone and those treated with methotrexate and histidine. NOD-SCID mice were injected with HEL cells subcutaneously on day 1, followed by weight measurement every other day, in vivo imaging of HEL-cell-derived tumours (on days 7, 15, 20 and 28), treatment on days 12 to 23 (vehicle, histidine and methotrexate injections every other day), and final termination of the experiment after 40 days, unless an early euthanization was required in accordance with the guidelines for humane experimental end-point of the Committee on Animal Care at MIT. The experiment included four experimental groups: vehicle-treated (saline) ($n = 7$), histidine supplementation ($n = 8$), methotrexate-treated ($n = 8$), and histidine supplementation combined with methotrexate treatment ($n = 8$). Serum was collected at day 0 and day 23 for metabolite profiling and liver diagnostics. The dose of methotrexate used was 25 mg kg^{-1} based on the weight measured at day 0. The dose of histidine was 18 mg per injection in 400 μl saline. **b**, A significant reduction in tumour size over time was found in mice treated with the combination of methotrexate and histidine supplementation. Tumours were imaged in vivo by luciferase expression at the indicated days. Fold changes in tumour sizes over measurements performed on day 7 are presented. *P* values were calculated by non-parametric one-way ANOVA. Group size changed over time owing to mouse euthanization for humane reasons. On day 15: vehicle-treated ($n = 5$), histidine supplementation ($n = 7$), methotrexate-treated ($n = 7$) and histidine supplementation combined with methotrexate treatment ($n = 8$). On day 20: vehicle-treated ($n = 6$), histidine supplementation ($n = 6$), methotrexate-treated ($n = 7$) and histidine supplementation combined with methotrexate treatment ($n = 8$). On day 28: vehicle-treated ($n = 6$), histidine supplementation ($n = 5$), methotrexate-treated ($n = 7$) and histidine supplementation combined with methotrexate treatment ($n = 8$). **c**, In vivo imaging of luciferase-expressing HEL-cell-derived tumours at days 7 (top) and 28 (bottom). Mice are labelled by number, with the colour of the label representing their experimental group. All mice assigned to the experiment are shown, group size is the same as in **a**. **d**, **e**, There was no increase in the abundance of serum markers indicative of kidney damage in mice treated with methotrexate + histidine supplementation compared to mice treated with methotrexate. Markers of kidney toxicity (urea and creatinine) were measured by LC–MS in serum samples of the tested mice. The relative abundance of urea and creatinine was normalized to isotopically labelled valine and tryptophan as an internal standard. *P* values were calculated using non-parametric one-way ANOVA. Group size: vehicle-treated ($n = 7$), histidine supplementation ($n = 7$), methotrexate-treated ($n = 8$), and histidine supplementation combined with methotrexate treatment ($n = 6$). **f**, **g**, Some increase in liver-toxicity markers in mice treated with the combined therapy compared to mice treated with methotrexate alone. Markers of liver toxicity (alanine aminotransferase (ALT) and aspartate aminotransferase (AST)) were measured by an external serum diagnostics laboratory (IDEXX) in serum samples of the tested mice. Measurement units are indicated. *P* values were calculated using non-parametric one-way ANOVA. Group size: vehicle-treated ($n = 6$), histidine supplementation ($n = 6$), methotrexate-treated ($n = 6$), and histidine supplementation combined with methotrexate treatment ($n = 5$). Centre bars indicate mean, error bars indicate s.e.m.

© 2018 Springer Nature Limited. All rights reserved.





Extended Data Fig. 11 | In vivo histidine supplementation sensitizes tumours to methotrexate without enhancement of treatment toxicity (part 2). a, **b**, Histological analyses indicated that the kidney and liver appeared normal at the end of the two-week recovery period. See also Supplementary Information Fig. 3a. **a**, H&E analyses of kidney sections from mice from all experimental groups. Tissues were collected at the conclusion of the experiment, after two weeks post-treatment recovery. Group sizes are the same as in Extended Data Fig. 9a. **b**, H&E analyses of liver sections from mice from all experimental groups. Group sizes are the same as in Extended Data Fig. 9a. Tissues were collected at the conclusion of the experiment, after two weeks post-treatment recovery. See also Supplementary Information Fig. 3b.

© 2018 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0350-5

Glucose–regulated phosphorylation of TET2 by AMPK reveals a pathway linking diabetes to cancer

Di Wu[1,2,7], Di Hu[1,2,7], Hao Chen[1,3,7], Guoming Shi[1,2,7], Irfete S. Fetahu[2], Feizhen Wu[1,2], Kimberlie Rabidou[2], Rui Fang[2], Li Tan[1], Shuyun Xu[4], Hang Liu[1], Christian Argueta[2], Lei Zhang[5], Fei Mao[6], Guoquan Yan[5], Jiajia Chen[5], Zhaoru Dong[1], Ruitu Lv[1], Yufei Xu[2], Mei Wang[2], Yong Ye[1], Shike Zhang[2], Danielle Duquette[2], Songmei Geng[4], Clark Yin[2], Christine Guo Lian[4], George F. Murphy[4], Gail K. Adler[2], Rajesh Garg[2], Lydia Lynch[2], Pengyuan Yang[5], Yiming Li[6], Fei Lan[1], Jia Fan[1], Yang Shi[1,3] & Yujiang Geno Shi[1,2]*

Diabetes is a complex metabolic syndrome that is characterized by prolonged high blood glucose levels and frequently associated with life-threatening complications[1,2]. Epidemiological studies have suggested that diabetes is also linked to an increased risk of cancer[3–5]. High glucose levels may be a prevailing factor that contributes to the link between diabetes and cancer, but little is known about the molecular basis of this link and how the high glucose state may drive genetic and/or epigenetic alterations that result in a cancer phenotype. Here we show that hyperglycaemic conditions have an adverse effect on the DNA 5-hydroxymethylome. We identify the tumour suppressor TET2 as a substrate of the AMP-activated kinase (AMPK), which phosphorylates TET2 at serine 99, thereby stabilizing the tumour suppressor. Increased glucose levels impede AMPK-mediated phosphorylation at serine 99, which results in the destabilization of TET2 followed by dysregulation of both 5-hydroxymethylcytosine (5hmC) and the tumour suppressive function of TET2 in vitro and in vivo. Treatment with the anti-diabetic drug metformin protects AMPK-mediated phosphorylation of serine 99, thereby increasing TET2 stability and 5hmC levels. These findings define a novel 'phospho-switch' that regulates TET2 stability and a regulatory pathway that links glucose and AMPK to TET2 and 5hmC, which connects diabetes to cancer. Our data also unravel an epigenetic pathway by which metformin mediates tumour suppression. Thus, this study presents a new model for how a pernicious environment can directly reprogram the epigenome towards an oncogenic state, offering a potential strategy for cancer prevention and treatment.

DNA methylation (5mC) and hydroxymethylation (5hmC) are epigenetic modifications that are frequently perturbed in cancer[6–8]. The conversion of 5mC to 5hmC occurs through an oxidative reaction catalysed by the ten-eleven translocation (TET) protein family of dioxygenases (TET1, TET2 and TET3)[9–11]. The reaction requires α-ketoglutarate, a metabolite that is influenced by the availability of glucose and glutamine[12,13]. This led us to predict that a hyperglycaemic state would increase the levels of 5hmC in blood cells from patients with diabetes. To test this possibility, we examined global 5hmC in gDNA extracted from peripheral blood mononuclear cells (PBMC) collected from a cohort of healthy donors (haemoglobin A1c (HbA1c) 5.5 ± 0.26%) and patients with diabetes (HbA1c 10.7 ± 1.9%, Fig. 1a). Unexpectedly, samples from patients showed a significant ($P = 0.0017$) decrease in 5hmC levels compared to the healthy donors (Fig. 1b, Extended Data Fig. 1a), whereas 5mC levels remained the same (Fig. 1c). The presence of diabetes in the patient group was the strongest predictor of low 5hmC levels, with HbA1c levels showing a significant ($P < 0.05$) inverse correlation with 5hmC (Extended Data Table 1a).

To investigate this inverse correlation, we cultured several cell lines under normal glucose (1 g l^{-1}) and high glucose (4.5 g l^{-1}) conditions. A subset of these cell lines (PBMC, HUVEC and TF-1) exhibited significantly ($P = 0.022$ (PBMC), 0.046 (HUVEC), 0.047 (TF-1)) lower levels of 5hmC when subjected to high as opposed to normal glucose. The other cell lines (A375, A2058 and SK-MEL-5) did not show apparent changes in 5hmC levels between the two glucose conditions, as they have low baseline levels of 5hmC (Extended Data Fig. 1b).

Loss of 5hmC is an epigenetic hallmark of cancer, in which diminished levels of TET2 expression have an important role[14,15]. We hypothesized that the alterations in 5hmC in response to glucose were mediated through TET2, as TET1 and TET3 were barely detectable in these cells. Indeed, glucose-responsive cells (PBMC, HUVEC and TF-1) showed decreased TET2 in high glucose, whereas TET2 remained low and unaltered in the glucose-nonresponsive cells (A2058, A375 and SK-MEL-5) (Extended Data Fig. 1c). To unravel the specific role of TET2 in this modulation, we used an A2058-TET2WT stable cell line, in which the expression of TET1 and TET3 is low and the expression of TET2 and 5hmC levels have been restored[15] (Extended Data Fig. 1d, e). High glucose reduced the level of TET2 protein in TET2WT cells, whereas the *TET2* mRNA level remained unchanged (Extended Data Fig. 1f, g). Notably, the protein half-life of TET2 was substantially lower under high glucose than under normal glucose (Extended Data Fig. 1h). Unlike native A2058 cells (Extended Data Fig. 1b), A2058-TET2WT cells showed a significant ($P = 0.034$) increase in 5hmC levels after 7 days of culture in normal glucose. This increase, however, could be reversed (within 24 h) by switching the culturing medium to high glucose (Fig. 1d, Extended Data Fig. 1i, j; see Supplementary Information for details). Such modulation was not observed in stable A2058 cell lines expressing mock control (mock), catalytically inactive full-length TET2 (TET2M), or catalytically active C-terminal TET2 (TET2CD) (Fig. 1e, Extended Data Fig. 1k). These data suggest that functional, full-length TET2 is required to mediate the reversible changes in 5hmC in response to extracellular glucose availability.

To identify gene-specific 5hmC changes in response to glucose, we analysed genome-wide alterations in 5hmC in A2058-TET2WT cells using hydroxymethylated DNA immunoprecipitation coupled with deep sequencing (hMeDIP–seq) and methylated DNA immunoprecipitation coupled with deep sequencing (MeDIP–seq). These analyses showed that 5hmC was higher under normal glucose than high glucose, whereas there was no significant difference in 5mC (Fig. 1f, Extended Data Fig. 2a). We identified 30,217 differentially 5-hydroxymethylated regions (DhMRs), with more than 80% (about 24,537) of the regions being increased under normal glucose (Extended Data Fig. 2b, c). The majority of this DhMR enrichment (65.4%) was associated with

[1]Key Laboratory of Medical Epigenetics and Metabolism, Institute of Clinical Science of Zhongshan Hospital and Institutes of Biomedical Sciences, Fudan University, Shanghai, China. [2]Division of Endocrinology, Diabetes and Hypertension, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [3]Division of Newborn Medicine and Program in Epigenetics, Department of Medicine, Boston Children's Hospital, Boston, MA, USA. [4]Program in Dermatopathology, Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [5]Department of Chemistry and Institutes of Biomedical Science, Shanghai Medical School, Fudan University, Shanghai, China. [6]Division of Endocrinology and Metabolism, Huashan Hospital, Fudan University, Shanghai, China. [7]These authors contributed equally: Di Wu, Di Hu, Hao Chen, Guoming Shi. *e-mail: yujiang_shi@hms.harvard.edu

© 2018 Springer Nature Limited. All rights reserved.

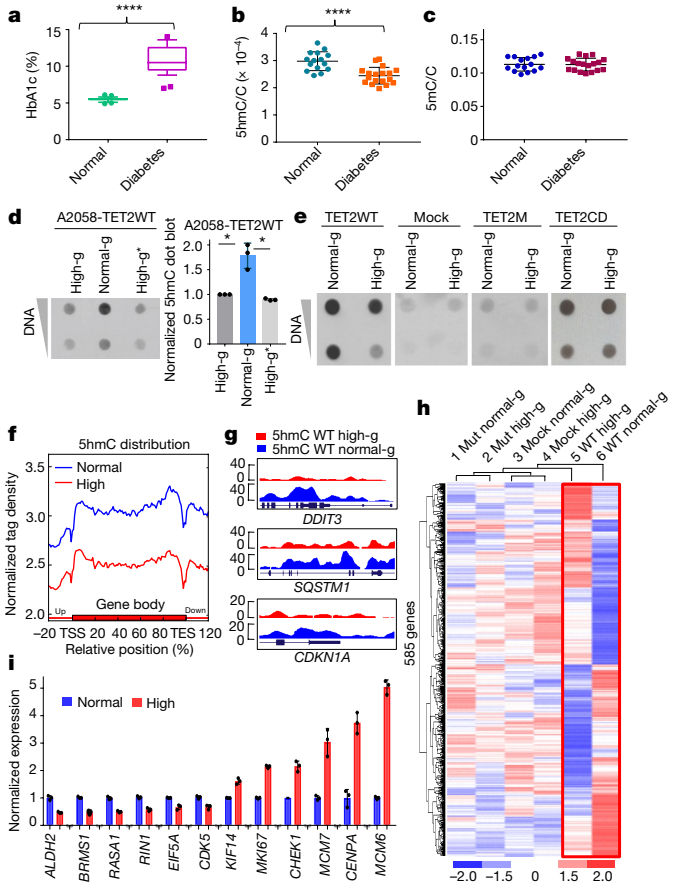


Fig. 1 | Hyperglycaemia deregulates global DNA 5hmC levels, and this regulation requires a functional, full-length TET2. a, Comparison of HbA1c between 28 healthy donors and 29 patients with diabetes. $^{****}P = 2.88 \times 10^{-15}$. Box plot: centre lines, median; limits, upper and lower quartiles; top and bottom whiskers, 10% and 90% percentiles; points, outliers. **b**, High-performance liquid chromatography with tandem mass spectrometry (HPLC–MS/MS) analysis of ratio of 5hmC to total C in randomly selected gDNA from 15 healthy donors and 18 patients. $^{****}P = 7.44 \times 10^{-5}$. **c**, HPLC–MS/MS analysis of ratio of 5mC to total C from the same samples as in **b**. **d**, Left, dot blot of 5hmC in DNA extracted from A2058-TET2WT cell line cultured in high glucose (high-g) or normal glucose (normal-g), or switched from normal to high glucose (high-g*). Right, quantification of dot blots. $^{*}P = 0.034$ (high-g), $P = 0.026$ (high-g*). **e**, Dot blot comparison of 5hmC levels in A2058-TET2WT, mock (an A2058 cell line stably expressing empty vector), A2058-TET2M, and A2058-TET2CD cell lines cultured in high and normal glucose. **f**, Normalized hMeDIP–seq 5hmC tag density distribution across gene bodies in A2058-TET2WT cells cultured in high glucose (red) or normal glucose (blue). Each gene body was normalized relative to position percentage within the gene. **g**, hMeDIP–seq results of representative genes in which 5hmC is increased across the gene body in normal glucose. *y*-axis represents 5hmC density. **h**, Unguided clustering of genes that are differentially expressed when TET2 is present (in red frame) and not differentially expressed when TET2 is absent. Mut, A2058-TET2M. **i**, RT–qPCR validation of a subset of cancer-associated genes identified in **h**. Data represent three technical repeats in **i** and three biologically independent repeats in **d**, **e**. Two-sided Student's *t*-test, data shown as mean ± s.d., $^{*}P < 0.05$, $^{****}P < 0.0001$.

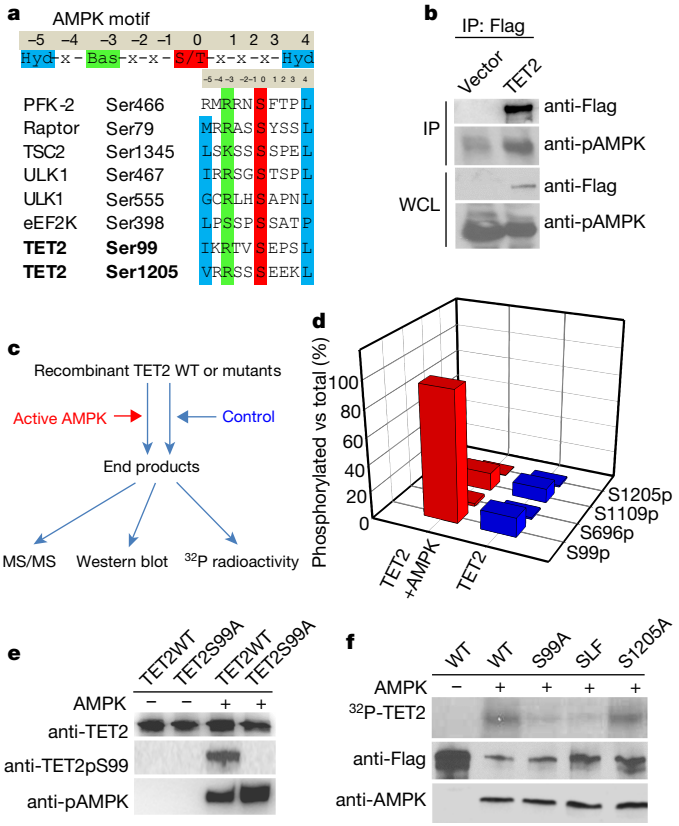


Fig. 2 | TET2 is a substrate of AMPK, which specifically phosphorylates TET2 at S99. a, Conserved AMPK recognition motif of several key AMPK target enzymes, including TET2. **b**, Co-immunoprecipitation showing that Flag–TET2 interacts with the activated form of AMPK (AMPKpT172). **c**, Flowchart of the in vitro kinase assay with different detection approaches. **d**, LC–MS/MS profiling of the in vitro AMPK kinase assay showed that TET2S99 is specifically phosphorylated in the presence of active AMPK. **e**, Immunoblot validation of TET2pS99 phosphorylation using a TET2pS99-specific antibody. **f**, Determination of TET2 phosphorylation at S99 via ^{32}P-autoradiograph (SLF is a naturally occurring mutant that probably disrupts the AMPK binding site). Data represent three biologically independent repeats in **b**, **e**, **f**.

gene regions, with 9.56% localizing to promoters and 55.84% to gene bodies (Fig. 1g, Extended Data Fig. 2d). Notably, gene ontology and disease ontology analyses of these DhMR enriched genes showed strong associations with cancer and cancer-related pathways (Extended Data Fig. 2e).

We next used microarrays to compare transcriptomes of mock, A2058-TET2M and A2058-TET2WT cells cultured in normal or high glucose (Fig. 1h). This analysis identified 585 genes that were differentially expressed in normal glucose versus high glucose when TET2 was present (A2058-TET2WT), but not in mock or A2058-TET2M cells (Fig. 1h, Extended Data Table 1b). These glucose-modulated and TET2-dependent genes are involved in cell cycle regulation and are associated with various cancers (Extended Data Fig. 3a, b). Among these genes, we found several tumour suppressor and tumour promoting genes that were downregulated and upregulated, respectively, under the high glucose condition, and validated a subset of these using quantitative PCR with reverse transcription (RT–qPCR) (Fig. 1i). Furthermore, we found that 213 of the 585 genes (36.4%) contained increased DhMRs in normal glucose. Gene ontology and disease ontology analysis revealed that these 213 genes were also highly associated with cell cycle regulation and various types of cancers (Extended Data Fig. 3c). Of note, these genes consist of both upregulated and down-regulated genes (Extended Data Fig. 3d), suggesting that 5hmC has a complex role in gene regulation, as previously reported[16,17].

AMPK is a key nutrient or energy sensor that is highly sensitive to glucose availability[18]. TET2 contains two putative AMPK motifs centred on serines 99 (S99) and 1205 (S1205) (Fig. 2a). Furthermore, the activated form of AMPK (pAMPK172) was co-immunoprecipitated with TET2 (Fig. 2b). To show that TET2 is a substrate of AMPK, we carried out an in vitro AMPK kinase assay (Fig. 2c, Extended Data Fig. 4a). By liquid chromatography with tandam mass spectrometry (LC–MS/MS) analysis, we detected robust phosphorylation of TET2, specifically at S99, by the active form of AMPK (Fig. 2d, Extended Data

© 2018 Springer Nature Limited. All rights reserved.

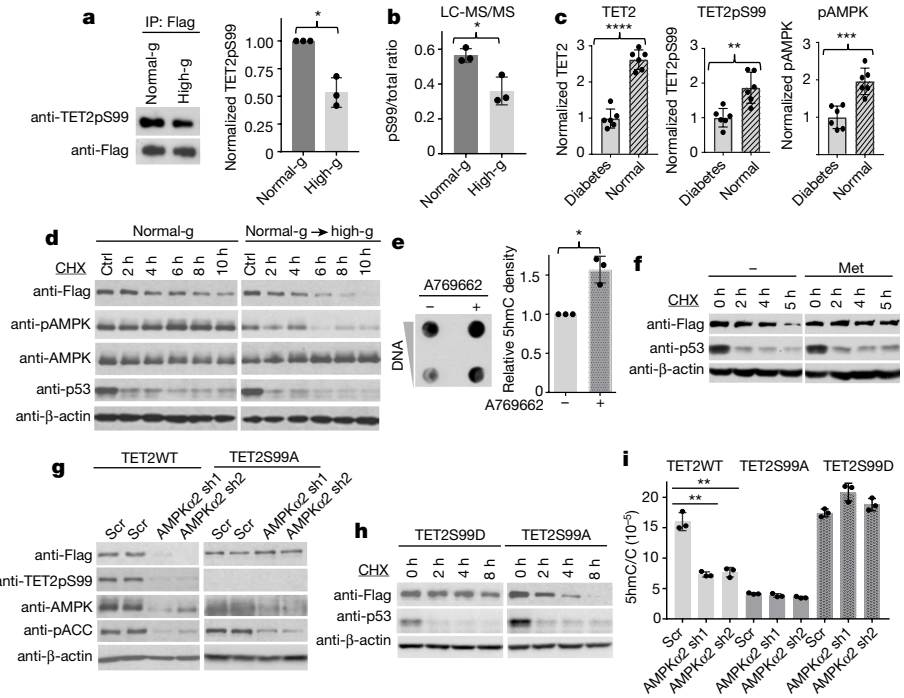


Fig. 3 | Glucose regulates TET2 stability via the TET2pS99 phosphorylation switch, which is controlled by AMPK. a, Left, immunoblot signal of phosphoserine 99 (TET2pS99) on Flag–TET2WT immunoprecipitated from cells cultured in high or normal glucose. Right, quantification of TET2pS99 signals normalized by total TET2, $*P = 0.027$. **b**, LC–MS/MS validation of pS99 levels in the cells under high and normal glucose, $*P = 0.030$. **c**, Normalized TET2, TET2pS99 and pAMPK levels in PBMCs from healthy donors and patients with diabetes; $n = 6$, $****P = 1.01 \times 10^{-6}$ (TET2), $**P = 0.0035$ (TET2pS99), $***P = 5.26 \times 10^{-4}$ (pAMPK). **d**, Flag–TET2WT half-life and pAMPK levels under normal glucose or normal glucose switched to high glucose. **e**, Left, dot blot showing 5hmC levels in A2058-TET2WT cells after treatment with 100 μM A769662 (an AMPK activator). Right, normalized quantification of dot blot signals; $*P = 0.029$. **f**, CHX treatment to measure the half-life of Flag–TET2WT in cells treated with (+) or without (−) metformin (Met; 5 mM). **g**, Effects of knocking down AMPKα2 on the protein stabilities of Flag–TET2WT and TET2S99A. Scr, scrambled shRNA. **h**, Half-lives of Flag–TET2S99D and TET2S99A. **i**, HPLC–MS/MS results showing the effects of knocking down AMPKα2 on 5hmC levels in TET2WT ($**P = 0.006$ (sh1), 0.003 (sh2)) cells and cells overexpressing TET2S99A or TET2S99D. Data represent three biologically independent repeats in **a**, **b**, **d**–**i**. Two-sided Student's *t*-test, data shown as mean ± s.d., $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, $****P < 0.0001$. For gel source data, see Supplementary Fig. 1.

Fig. 4b). There was little change to the other putative phosphorylation sites, including S1205. We next generated an antibody that specifically recognizes the phosphorylated form of TET2 S99 (Extended Data Fig. 4c). Immunoblotting confirmed the specificity of the antibody, as it gave a strong signal when used against AMPK-treated TET2WT, but not with untreated TET2WT or TET2 with a serine-to-alanine mutation at position 99 (TET2S99A) (Fig. 2e). Using a ^{32}P radiolabelling kinase assay, we observed a marked decrease in γ-ATP incorporation in the TET2S99A mutant as well as in the S102_L103 >F (SLF) mutant (http://cancer.sanger.ac.uk/cosmic). By contrast, TET2WT and the S1205A mutant demonstrated comparably high levels of radioactivity (Fig. 2f, Extended Data Fig. 4d). These findings reveal that TET2 is a substrate of AMPK and that S99 is the major phosphorylation site catalysed by AMPK.

AMPK is known to be activated upon energy stress induced by glucose depletion[19,20]. We hypothesized that glucose-mediated TET2 protein stability was modulated through phosphorylation of TET2S99 by AMPK. To test this hypothesis, we first compared TET2S99 phosphorylation levels in A2058-TET2WT cells cultured in normal glucose versus high glucose. Both western blot ($P = 0.027$) and HPLC–MS analysis ($P = 0.030$) revealed that the phosphorylation status of TET2S99 (TET2pS99) was significantly higher in normal as compared to high glucose (Fig. 3a, b, Extended Data Fig. 5a). Consistent with the kinetics of 5hmC shown in Extended Data Fig. 1i, j, we observed steady increases in pAMPK, TET2pS99 and TET2 during days 0, 2 and 4 after switching from high to normal glucose (Extended Data Fig. 5b). These increases were confirmed in other cell lines as well as in PBMCs (Extended Data Fig. 5c). Notably, pAMPK, TET2pS99 and total TET2 were significantly ($P = 1.01 \times 10^{-6}$ (TET2), 0.0035 (TET2pS99), 5.26×10^{-4} (pAMPK)) higher in PBMCs from healthy donors than in those from patients with diabetes (Fig. 3c, Extended Data Fig. 5d). This is consistent with our observation that 5hmC is increased in the healthy group (Fig. 1a, b). Collectively, these data demonstrate that TET2 is phosphorylated at S99 by AMPK, which is suppressed under hyperglycaemic conditions.

To investigate the pivotal role of AMPK-mediated phosphorylation in TET2 stability and 5hmC levels, we measured the half-life of TET2 and corresponding AMPK activity in cells that were initially cultured in normal glucose and were then switched to high glucose or kept in normal glucose. As indicated by pAMPK (pThr172), AMPK retained phosphorylation activity in cells that were continually cultured in normal glucose, whereas it was deactivated in cells switched to high glucose. Concomitantly, we observed that the half-life of TET2 protein was substantially reduced in cells that were switched to high glucose (Fig. 3d). Next, we treated cells with AMPK activators (metformin or A769662)[21,22]. As predicted, the activators elevated the levels of pAMPK, TET2pS99, total TET2 (Extended Data Fig. 6a, b) and 5hmC (Fig. 3e). When cells cultured in high glucose were treated with metformin or A769662, the half-life of TET2 increased notably (Fig. 3f, Extended Data Fig. 6c, d). Furthermore, metformin treatment also increased 5hmC in a subset of DhMRs previously identified to be increased under normal glucose (Extended Data Fig. 6e). To demonstrate that AMPK directly regulates TET2 protein stability, we used short hairpin RNAs (shRNAs) to specifically knockdown AMPKα2, an AMPKα isoform enriched in the nucleus[23], in A2058 cells expressing TET2WT or TET2S99A, the non-phosphorylatable mutant. *AMPKα2* shRNAs effectively depleted AMPKα2 and resulted in a marked decrease in TET2WT protein but not in the S99A mutant

© 2018 Springer Nature Limited. All rights reserved.

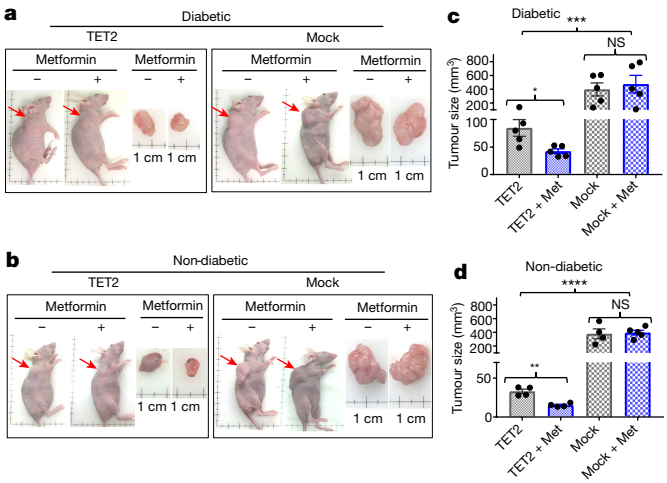


Fig. 4 | Hyperglycaemia promotes xenografted tumour growth, which is suppressed by TET2 and further prevented by metformin. **a**, **b**, Representative images of tumour-bearing diabetic (**a**) and non-diabetic (**b**) nude mice xenografted with A2058-TET2WT (left) or mock tumour cells (right) treated with (+) or without (−) metformin. **c**, **d**, Quantitative comparison of tumour size and efficacy of metformin's tumour suppressive function on A2058-TET2WT and mock tumours in diabetic (**c**) and non-diabetic (**d**) nude mice. **c**, $n = 5$, $^{***}P = 8.3 \times 10^{-4}$ (TET2 versus mock), $^{*}P = 0.048$ (TET2 versus TET2 + Met). **d**, $n = 4$–5, $^{****}P = 5.2 \times 10^{-6}$ (TET2 versus mock), 0.0046 (TET2 versus TET2 + Met). Two-sided Student's *t*-test, data shown as mean ± s.e.m. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$; NS, not significant.

(Fig. 3g). These data suggest that AMPK and TET2S99 phosphorylation are important for protecting TET2 stability.

To identify the role of S99 phosphorylation, we generated another A2058 stable cell line expressing a phospho-mimic mutant, TET2S99D (Extended Data Fig. 7a, b).When treated with cycloheximide (CHX), the TET2S99A mutant exhibited less stability than the TET2S99D mutant (Fig. 3h). Consistent with greater protein stability, the TET2S99D cells had higher 5hmC than TET2S99A cells in high glucose (Fig. 3i). After knocking down AMPK, we observed a marked drop in 5hmC levels in wild-type cells, with no such effect in the TET2S99A or S99D cells (Fig. 3i, Extended Data Fig. 7c).

Previous studies have shown that the stability of TET2 protein is regulated by calpain family proteases[24], and phosphorylation can protect proteins from calpain cleavage[25]. Thus, we hypothesized that phosphorylation of S99 protects TET2 from calpain-mediated degradation. Indeed, treatment with a calpain inhibitor strongly stabilized TET2S99A and increased the protein level 3.7-fold, whereas only marginal effects were seen with TET2WT and TET2S99D (Extended Data Fig. 7d). Together, these results indicate that TET2pS99 is a 'phospho-switch' that is regulated by AMPK and is critical for the regulation of the stability of TET2 and levels of 5hmC.

To investigate how this AMPK–TET2 axis translates sustained high glucose exposure into a cancer-prone phenotype, we investigated the effects of TET2 and glucose levels on cell proliferation. A2058-TET2WT cells showed significantly ($P = 5.68 \times 10^{-5}$) higher cell proliferation rates in high glucose than in normal glucose, whereas mock cells showed little effect (Extended Data Fig. 8a). Notably, this TET2-dependent and glucose-influenced growth phenotype was validated in an independent model, TF-1 cells[26] (Extended Data Fig. 8b–d). These data suggest that TET2 has an important role in glucose-modulated tumour cell growth.

There is growing evidence that metformin, a widely used anti-diabetic drug, is also a potential anti-cancer agent[27,28]. We investigated whether the effect of metformin on cell proliferation also involves the AMPK–TET2–5hmC axis. We treated A2058-TET2WT, A2058-TET2S99A and mock cells with metformin and detected a significant ($P = 0.026$) increase in 5hmC levels in TET2WT cells, but not in TET2S99A or mock cells (Extended Data Fig. 8e). Using MTS (3-(4,5-dimethylthiazol-2-yl)-5-(3-carboxymethoxyphenyl)-2-(4-sulfophenyl)-2H-tetrazolium) proliferation assays, we examined the growth of cells treated with metformin and observed that metformin significantly ($P = 7.18 \times 10^{-5}$) reduced the proliferation of TET2WT cells, but had no effect on TET2S99A or mock cells (Extended Data Fig. 8f).

Next, we analysed the anchorage-independent growth of TET2WT, TET2S99A and mock cells. Restoring TET2 expression reduced anchorage-independent growth of A2058 cells. However, this suppression was compromised in A2058 cells expressing TET2S99A (Extended Data Fig. 8g, h, 0 mM). This phenomenon could be recapitulated in an independent cell line, MDA-MB-231 (Extended Data Fig. 8i–l). Notably, metformin further reduced colony growth in a dose-dependent manner, adding another layer of suppression to A2058-TET2WT cells. By contrast, the TET2S99A mutant and mock cells did not show significant growth inhibition upon metformin treatment (Extended Data Fig. 8g, h, 2–4 mM). Collectively, these results demonstrate in vitro that metformin requires the AMPK–TET2–5hmC axis to execute its anti-tumour effects.

To validate our hypothesis in vivo, we first generated BALB/c nude mice xenografted with TET2WT or mock A2058 tumours under both diabetic and non-diabetic conditions, as outlined in Extended Data Fig. 9a, b. In support of the role of TET2 as a tumour suppressor, TET2WT tumours were significantly smaller than those of the mock control in both diabetic (Fig. 4a, c; $P = 8.3 \times 10^{-4}$) and non-diabetic mice (Fig. 4b, d; $P = 5.2 \times 10^{-6}$). Notably, TET2WT tumours in diabetic mice were significantly larger than those in non-diabetic mice (Extended Data Fig. 9e, f; $P = 0.026$ without metformin treatment, $P = 0.0023$ with metformin treatment). By contrast, there was little difference between mock tumours grown in diabetic or non-diabetic mice. These results are consistent with a mechanism in which a sustained diabetic or hyperglycaemic environment impairs TET2–5hmC-mediated tumour suppression.

Next, we investigated whether we could observe the anti-tumour effects of metformin, also operating through the AMPK–TET2–5hmC axis, in vivo. We found that metformin imposed an additional layer of suppression on TET2WT tumours in both diabetic and non-diabetic mice (Extended Data Fig. 9c, d). By contrast, mock tumours showed little, if any, difference in size under the same treatments. Notably, when we depleted TET2 expression in A2058-TET2WT cells, the tumours behaved similarly to mock tumours and were no longer suppressed by metformin (Extended Data Fig. 9g, h). We should note that metformin did not change blood glucose levels, consistent with a previous report[29] (Extended Data Table 1c). Nevertheless, immunohistochemical (IHC) staining for pAMPK showed that metformin treatment effectively increased pAMPK in both TET2WT and mock tumours (Extended Data Fig. 10a, c). The levels of 5hmC, however, were increased only in TET2WT tumours treated with metformin (Extended Data Fig. 10b, d). These data demonstrate in vivo that the anti-tumour effect of metformin requires a functional TET2 protein and acts downstream of the influence of glucose upon the AMPK–TET2–5hmC axis.

In summary, we have shown that sustained hyperglycaemia destabilizes the tumour suppressor TET2 and deregulates levels of 5hmC. We describe an environment-to-epigenome signalling pathway, the glucose–AMPK–TET2–5hmC axis, which links the level of extracellular glucose to the dynamic regulation of 5hmC—and, ultimately, diabetes to cancer. We have identified AMPK-mediated TET2 phosphorylation at S99 as a molecular switch that controls a pivotal step in the glucose–AMPK–TET2–5hmC axis. Disabling this switch causes calpain-mediated degradation of TET2, resulting in a dysregulated hydroxymethylome and transcriptome. Notably, the anti-tumour effect of metformin requires a functional AMPK–TET2–5hmC axis. We propose that the glucose–AMPK–TET2–5hmC axis, which may work separately or in conjunction with other cellular pathways such as those involving mTOR[30], is a signalling pathway by which cancer-promoting environmental cues are directly linked to

© 2018 Springer Nature Limited. All rights reserved.

the reprogramming of a cancer-favourable epigenome. This epigenetic regulation may have a direct effect on the efficacy of metformin in preventing cancer, thus providing novel avenues for future clinical investigation.

Online content

Any Methods, including any statements of data availability and Nature Research reporting summaries, along with any additional references and Source Data files, are available in the online version of the paper at https://doi.org/10.1038/s41586-018-0350-5.

Received: 7 January 2016; Accepted: 7 June 2018;
Published online 18 July 2018.

1. Nathan, D. M. Long-term complications of diabetes mellitus. *N. Engl. J. Med.* **328**, 1676–1685 (1993).
2. Gregg, E. W. et al. Changes in diabetes-related complications in the United States, 1990–2010. *N. Engl. J. Med.* **370**, 1514–1523 (2014).
3. Giovannucci, E. et al. Diabetes and cancer: a consensus report. *CA Cancer J. Clin.* **60**, 207–221 (2010).
4. Chowdhury, T. A. Diabetes and cancer. *QJM* **103**, 905–915 (2010).
5. Brower, V. Illuminating the diabetes-cancer link. *J. Natl Cancer Inst.* **104**, 1048–1050 (2012).
6. Baylin, S. B. & Jones, P. A. A decade of exploring the cancer epigenome - biological and translational implications. *Nat. Rev. Cancer* **11**, 726–734 (2011).
7. Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* **13**, 484–492 (2012).
8. Ko, M., An, J. & Rao, A. DNA methylation and hydroxymethylation in hematologic differentiation and transformation. *Curr. Opin. Cell Biol.* **37**, 91–101 (2015).
9. Tahiliani, M. et al. Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. *Science* **324**, 930–935 (2009).
10. Zhang, H. et al. TET1 is a DNA-binding protein that modulates DNA methylation and gene transcription via hydroxylation of 5-methylcytosine. *Cell Res.* **20**, 1390–1393 (2010).
11. He, Y. F. et al. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**, 1303–1307 (2011).
12. Carey, B. W., Finley, L. W., Cross, J. R., Allis, C. D. & Thompson, C. B. Intracellular α-ketoglutarate maintains the pluripotency of embryonic stem cells. *Nature* **518**, 413–416 (2015).
13. Xu, W. et al. Oncometabolite 2-hydroxyglutarate is a competitive inhibitor of α-ketoglutarate-dependent dioxygenases. *Cancer Cell* **19**, 17–30 (2011).
14. Ko, M. et al. Impaired hydroxylation of 5-methylcytosine in myeloid cancers with mutant TET2. *Nature* **468**, 839–843 (2010).
15. Lian, C. G. et al. Loss of 5-hydroxymethylcytosine is an epigenetic hallmark of melanoma. *Cell* **150**, 1135–1146 (2012).
16. Williams, K. et al. TET1 and hydroxymethylcytosine in transcription and DNA methylation fidelity. *Nature* **473**, 343–348 (2011).
17. Xu, Y. et al. Genome-wide regulation of 5hmC, 5mC, and gene expression by Tet1 hydroxylase in mouse embryonic stem cells. *Mol. Cell* **42**, 451–464 (2011).
18. Hardie, D. G., Ross, F. A. & Hawley, S. A. AMPK: a nutrient and energy sensor that maintains energy homeostasis. *Nat. Rev. Mol. Cell Biol.* **13**, 251–262 (2012).
19. Hardie, D. G., Carling, D. & Carlson, M. The AMP-activated/SNF1 protein kinase subfamily: metabolic sensors of the eukaryotic cell? *Annu. Rev. Biochem.* **67**, 821–855 (1998).
20. de Lange, P. et al. Fuel economy in food-deprived skeletal muscle: signaling pathways and regulatory mechanisms. *FASEB J.* **21**, 3431–3441 (2007).
21. Zhou, G. et al. Role of AMP-activated protein kinase in mechanism of metformin action. *J. Clin. Invest.* **108**, 1167–1174 (2001).
22. Cool, B. et al. Identification and characterization of a small molecule AMPK activator that treats key components of type 2 diabetes and the metabolic syndrome. *Cell Metab.* **3**, 403–416 (2006).
23. Kim, N. et al. AMPKα2 translocates into the nucleus and interacts with hnRNP H: implications in metformin-mediated glucose uptake. *Cell. Signal.* **26**, 1800–1806 (2014).
24. Wang, Y. & Zhang, Y. Regulation of TET protein stability by calpains. *Cell Reports* **6**, 278–284 (2014).
25. Ma, S. et al. Site-specific phosphorylation protects glycogen synthase kinase-3β from calpain-mediated truncation of its N and C termini. *J. Biol. Chem.* **287**, 22521–22532 (2012).
26. Losman, J. A. et al. (R)-2-hydroxyglutarate is sufficient to promote leukemogenesis and its effects are reversible. *Science* **339**, 1621–1625 (2013).
27. Pernicova, I. & Korbonits, M. Metformin—mode of action and clinical implications for diabetes and cancer. *Nat. Rev. Endocrinol.* **10**, 143–156 (2014).
28. Morales, D. R. & Morris, A. D. Metformin in cancer treatment and prevention. *Annu. Rev. Med.* **66**, 17–29 (2015).
29. Sander, V. et al. Role of the N, N′-dimethylbiguanide metformin in the treatment of female prepuberal BALB/c mice hyperandrogenized with dehydroepiandrosterone. *Reproduction* **131**, 591–602 (2006).
30. Green, A. S. et al. LKB1/AMPK/mTOR signaling pathway in hematological malignancies: from metabolism to cancer cell biology. *Cell Cycle* **10**, 2115–2120 (2011).

Acknowledgements We thank A. Banks for discussions and reading of the manuscript; J. Cai for discussions and technical help; and L. Wang for technical support for PBMC culture. The TET1 antibody was a gift from G. Xu. This work was supported by NIH grants GM112062 and CA194302 to Y.G.S. See Supplementary Information for more details.

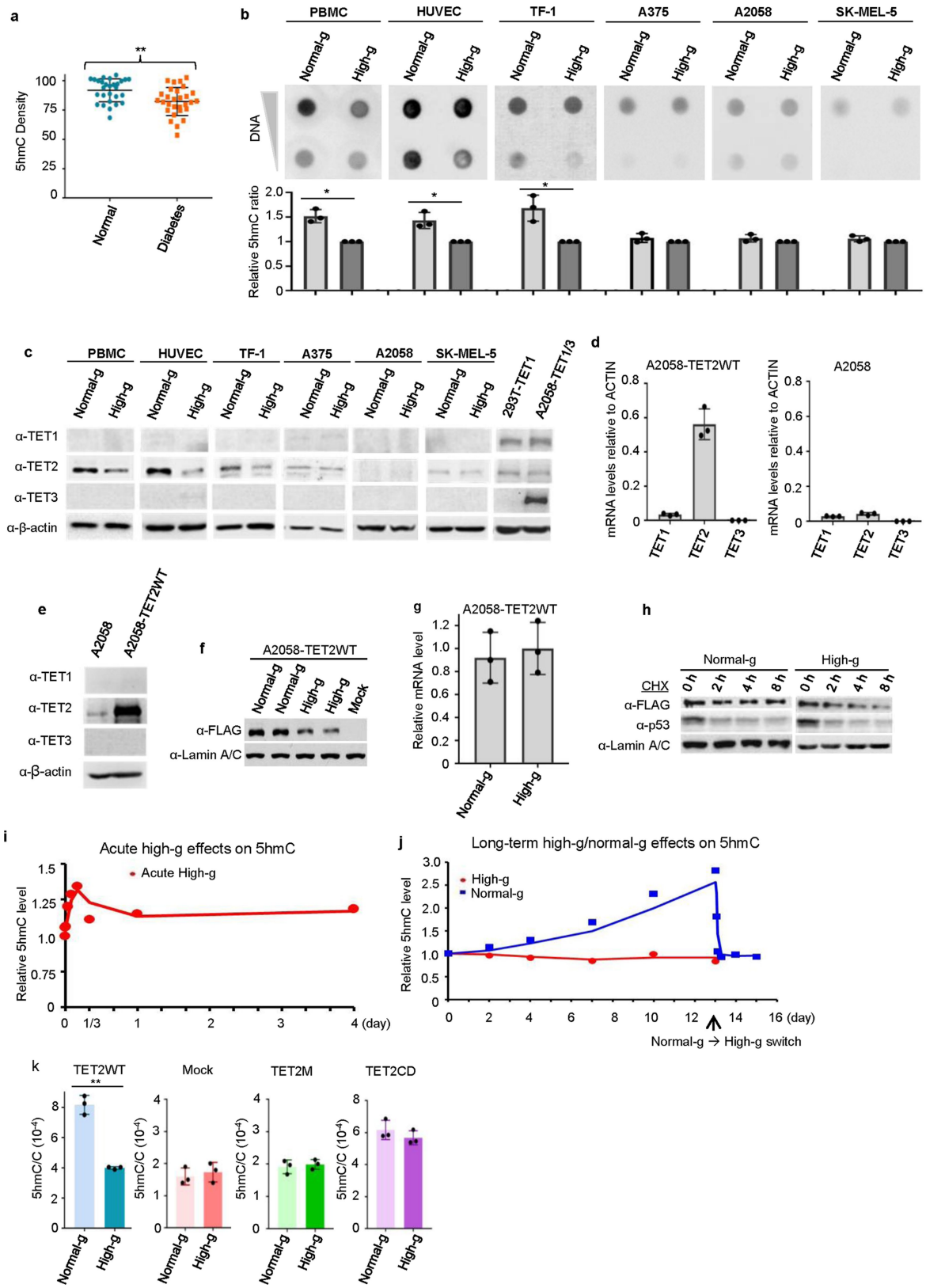
Reviewer information *Nature* thanks G. Hardie, R. Levine and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Y.G.S. conceived, designed and supervised the execution of the entire project. D.W. and D.H. initiated the project, designed experiments, and carried out major western blot, dot blot, patient sample, bioinformatic, cellular and animal studies. H.C. performed in vitro kinase assays and was responsible for making the TET2S99p antibody and TET2 mutant constructs. F.W., I.S.F., L.T. and H.L. were responsible for hMeDIP and sequencing analysis. R.F., C.A., C.Y., K.R. and S.Z. provided feedback on results. S.X. performed high-g/normal-g culturing and western blot for a number of cell lines. D.W. and R.L. were responsible for microarray analysis. L.Z., G.Y., J.C. and P.Y. were responsible for mass spectrometry analysis on DNA and protein. M.W., C.G.L. and G.F.M. performed experiments on dot blot. Y.Y., Z.D. and G.S. helped to establish animal studies. S.G. helped with IHC staining. Y.X. constructed TET1/2/3-overexpressing cell lines. F.M., D.D., G.K.A., R.G., L.L. and Y.L. coordinated work with diabetic patients. J.F., F.L., Y.L. and Y.S. provided constant ideas and discussions throughout the project and revised the manuscript. Y.G.S., D.W. and D.H. co-wrote the manuscript.

Competing interests Y.S. is cofounder of Constellation Pharmaceuticals, Inc and a member of its scientific advisory board. All other authors declare no competing financial interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0350-5.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0350-5.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to Y.G.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2018 Springer Nature Limited. All rights reserved.

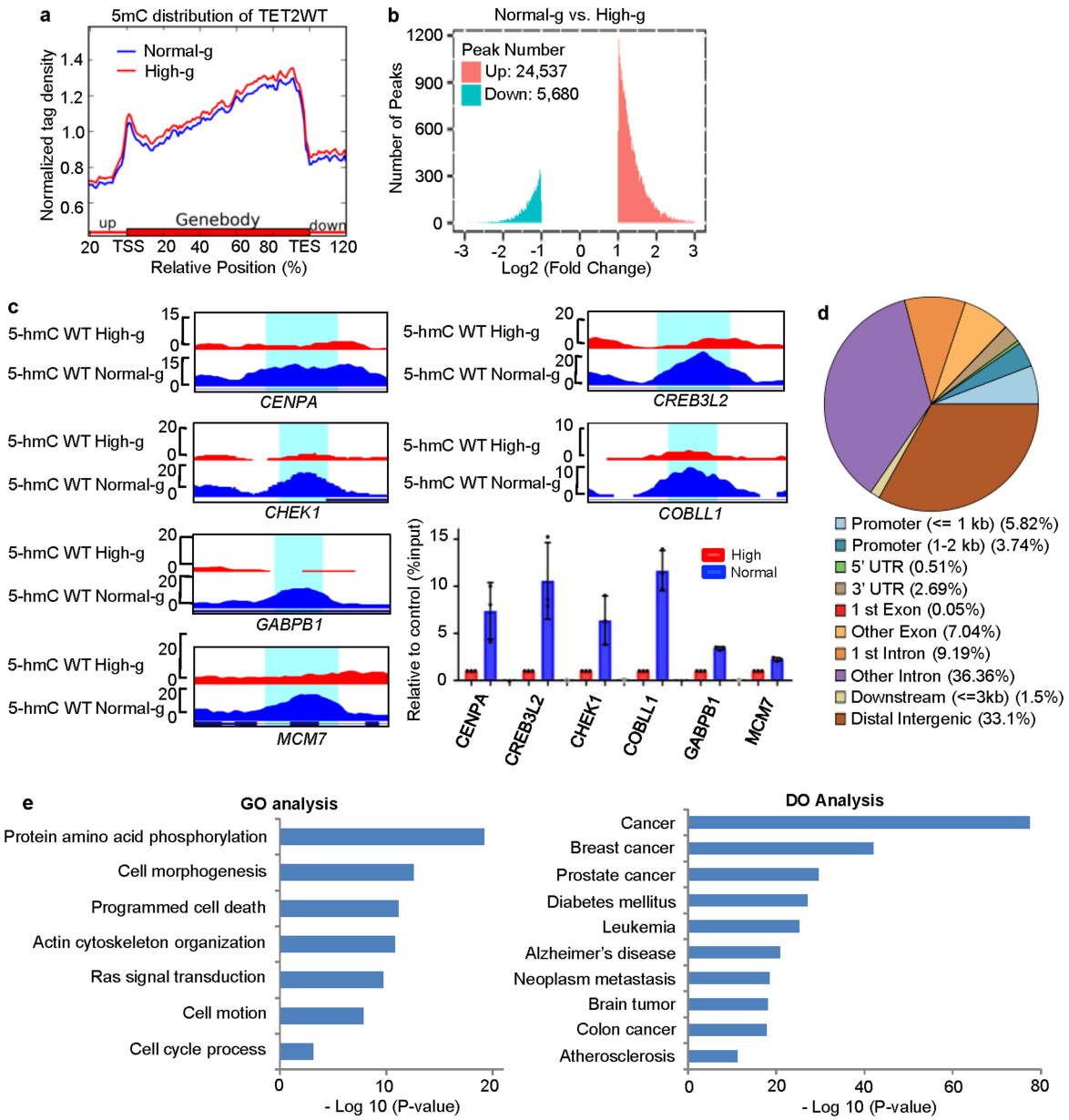


Extended Data Fig. 1 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

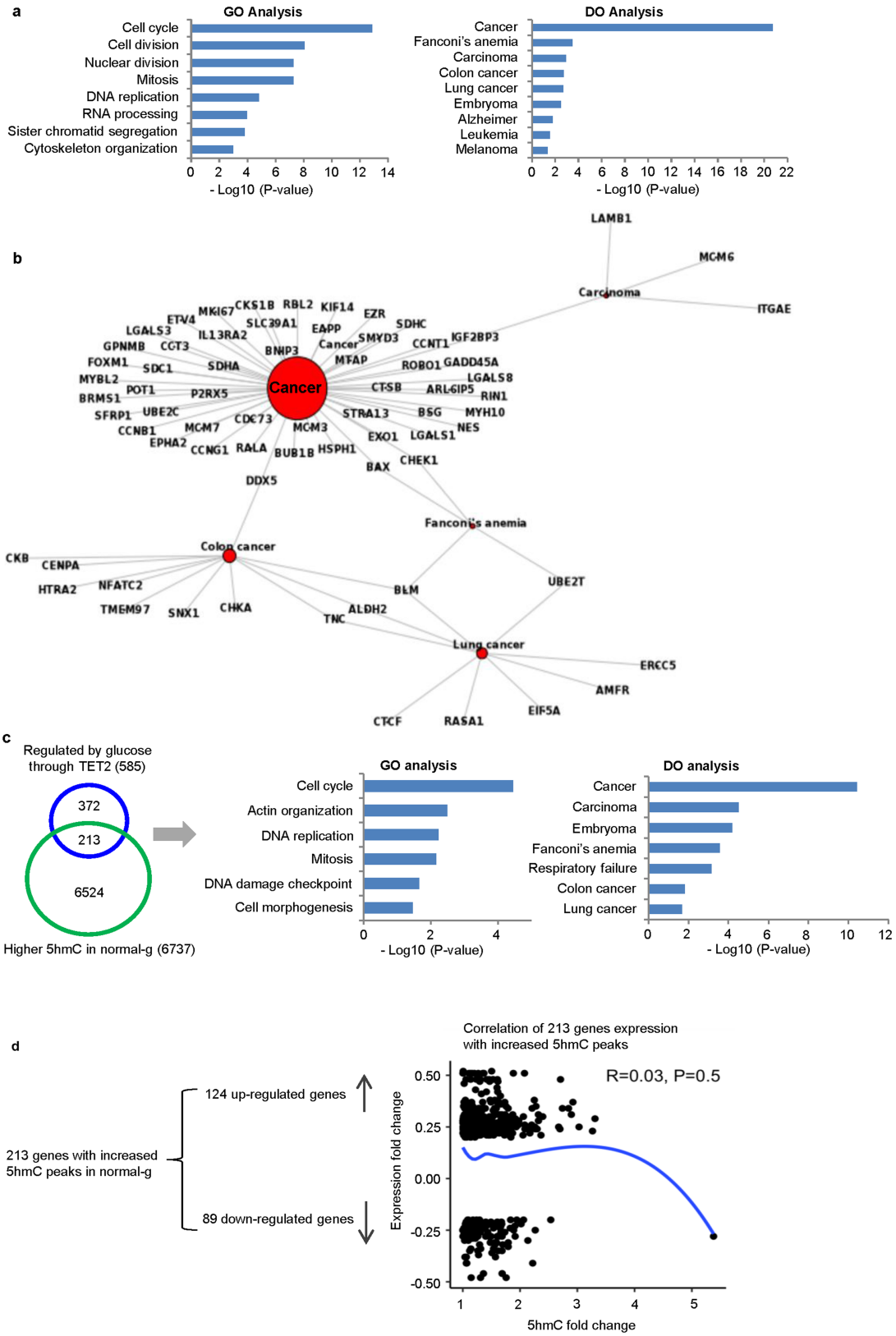
Extended Data Fig. 1 | Biochemical and molecular analysis of 5hmC and TET2 in response to glucose in primary blood cells and cultured cell lines. a, Dot blot comparison of global 5hmC levels between PBMC gDNA from 28 healthy donors and 29 patients with diabetes, $^{**}P = 0.0017$. **b**, Genomic DNA extracted from several cell types cultured under high or normal glucose were dot blotted for 5hmC, $^{*}P = 0.022$ (PBMC), 0.046 (HUVEC), 0.047 (TF-1). **c**, Western blot revealed that endogenous TET2 protein levels in PBMCs, HUVECs and TF-1 cells were higher when cultured in normal glucose than in high glucose. TET1 and TET3 protein levels were minimally detectable in these cell lines. **d**, **e**, RT–qPCR (**d**) and western blot (**e**) showed that TET2 dominated amongst the TET family in A2058-TET2WT cells. In A2058 cells, expression of all three genes was extremely low. **f**, Flag–TET2 protein levels in whole-cell lysates from A2058-TET2WT cells cultured in high glucose or normal glucose. **g**, Comparable levels of Flag–*TET2* mRNA from A2058-TET2WT cells cultured in high glucose or normal glucose. **h**, Half-lives of Flag–TET2 in A2058-TET2WT cells cultured in high glucose or normal glucose. **i**, High glucose shock resulted in an acute increase in 5hmC in A2058-TET2WT cells followed by a rapid drop to baseline. **j**, Long-term culturing of A2058-TET2WT cells in normal glucose resulted in a sustained increase in 5hmC, which could be reversed by switching the medium to high glucose. See Supplementary Information for more details. **k**, Dot blot quantification of 5hmC levels in A2058-TET2WT, mock, A2058-TET2M, and A2058-TET2CD cell lines in Fig. 1e; $^{**}P = 0.0068$. Data in **b**–**k** represent three biologically independent repeats each. Two-sided Student's *t*-test, data shown as mean ± s.d. $^{*}P < 0.05$, $^{**}P < 0.01$. For gel source data, see Supplementary Fig. 1.

© 2018 Springer Nature Limited. All rights reserved.



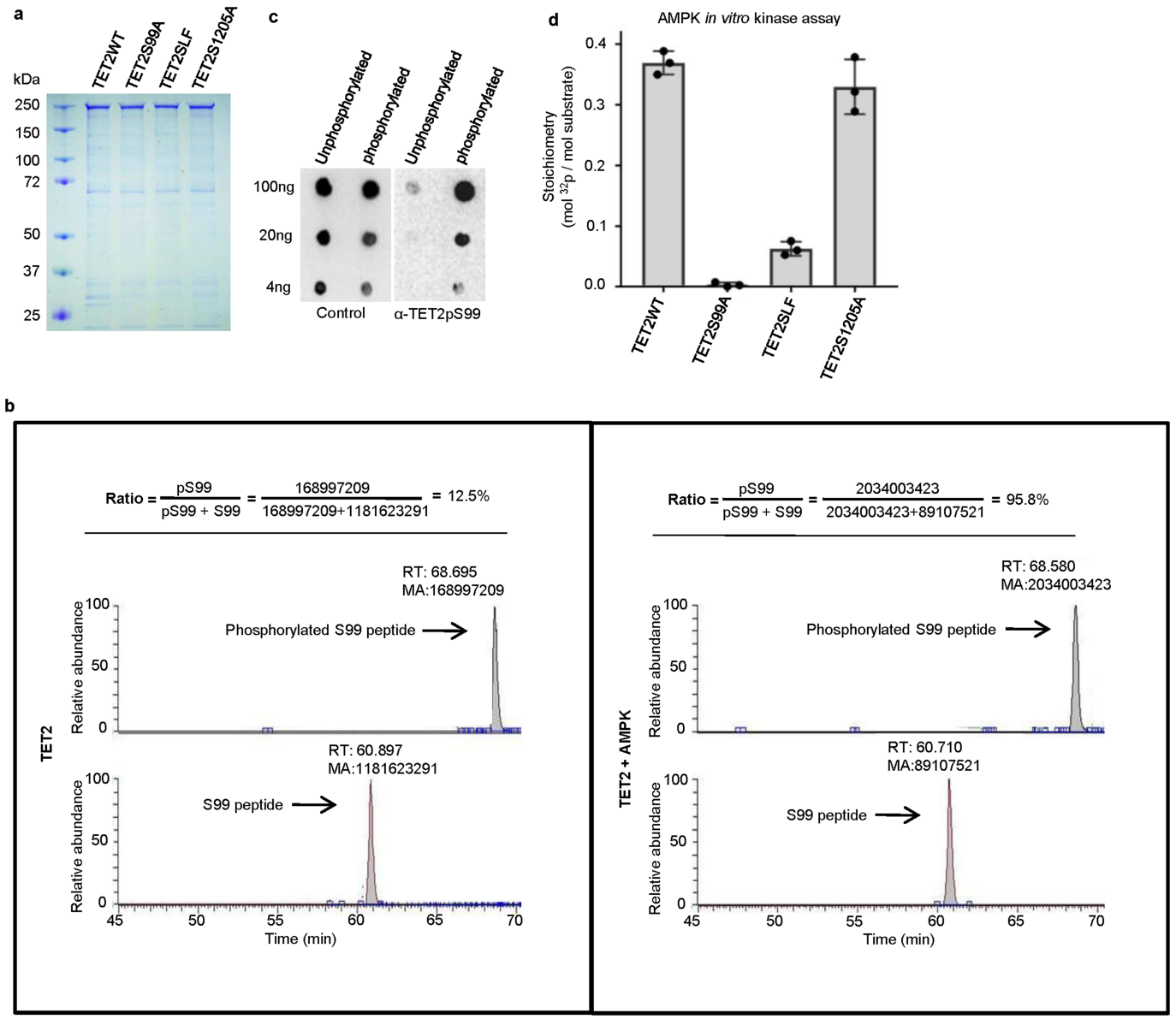
Extended Data Fig. 2 | Genome-wide alterations in 5mC and 5hmC under normal glucose compared with high glucose. a, Normalized 5mC tag density distribution of A2058-TET2WT across gene bodies under high glucose (red) and normal glucose (blue). Each gene body was normalized relative to position percentage within the gene. In contrast to 5hmC (Fig. 1f), the 5mC landscape of A2058-TET2WT cells showed marginal differences between high and normal glucose conditions. **b**, Total peak numbers for differentially 5-hydroxymethylated regions (DhMRs) when comparing normal glucose to high glucose. We identified 30,217 DhMRs in total, among which the majority (>80%, ~24,537) were increased in normal glucose, whereas <20% were decreased. **c**, hMeDIP–qPCR validation of DhMRs in a group of cancer-related genes including *CHEK1*, *CENPA*, *MCM7*, *CREB3L2*, *GABPB1* and *COBLL1* in A2058-TET2WT cells under high or normal glucose conditions; $n = 3$ biologically independent repeats. **d**, Genome-wide distribution of increased DhMRs in annotated genomic regions in A2058-TET2WT cells under normal glucose compared to high glucose. The majority of the DhMRs (65.4%) localized to transcribed gene regions, with 9.56% in promoters and 55.84% in gene bodies. **e**, Gene ontology (GO, left) and disease ontology (DO, right) analysis of genes ($n = 11,475$) that had increased DhMRs in A2058-TET2WT cells under normal glucose compared to high glucose. These genes were enriched in pathways related to cancer.

© 2018 Springer Nature Limited. All rights reserved.



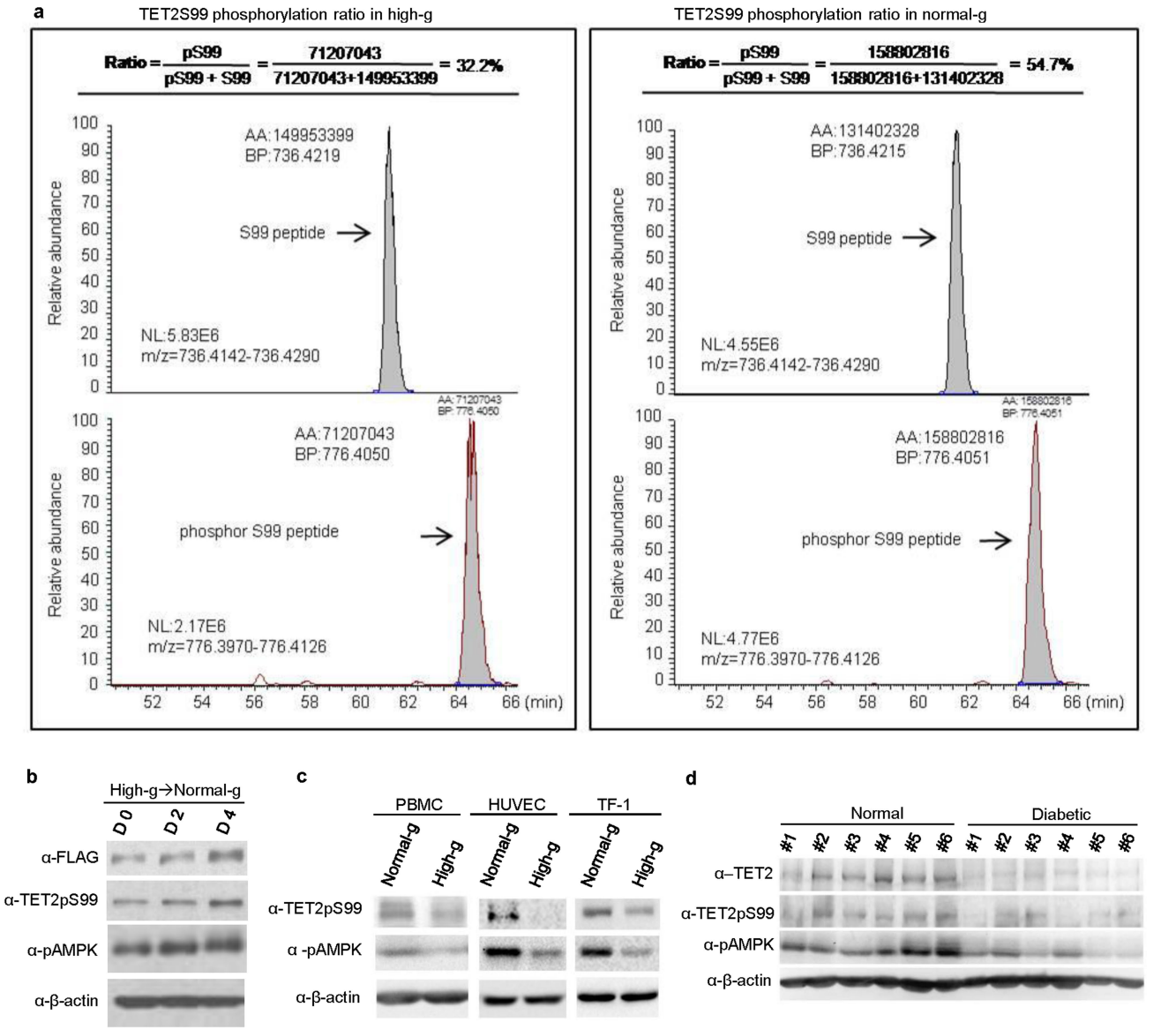
Extended Data Fig. 3 | Gene ontology and functional disease ontology analysis of genes that were transcriptionally regulated by glucose and TET2. a, The 585 glucose-modulated and TET2-dependent genes (Fig. 1h) identified in A2058-TET2WT cells were analysed by gene ontology (left) and disease ontology (right). These genes are enriched in cell cycle pathways and are strongly correlated with cancers, including carcinoma, colon cancer, lung cancer and embryoma. **b**, Detailed disease ontology clustering of the 585 genes showed strong association with various cancer types. **c**, Two hundred and thirteen out of the 585 genes also had increased DhMRs in normal glucose as compared to high glucose. Further analysis of this gene subset by gene ontology and disease ontology showed that these genes are also highly associated with cell cycle functions and strongly correlated with cancers. **d**, Of these 213 genes with increased 5hmC and altered gene expression in response to glucose alterations, 124 were upregulated and 89 were downregulated. Consistent with the intricate role of 5hmC in gene regulation, correlation analysis revealed no significant relationship between increased 5hmC and gene upregulation.

© 2018 Springer Nature Limited. All rights reserved.



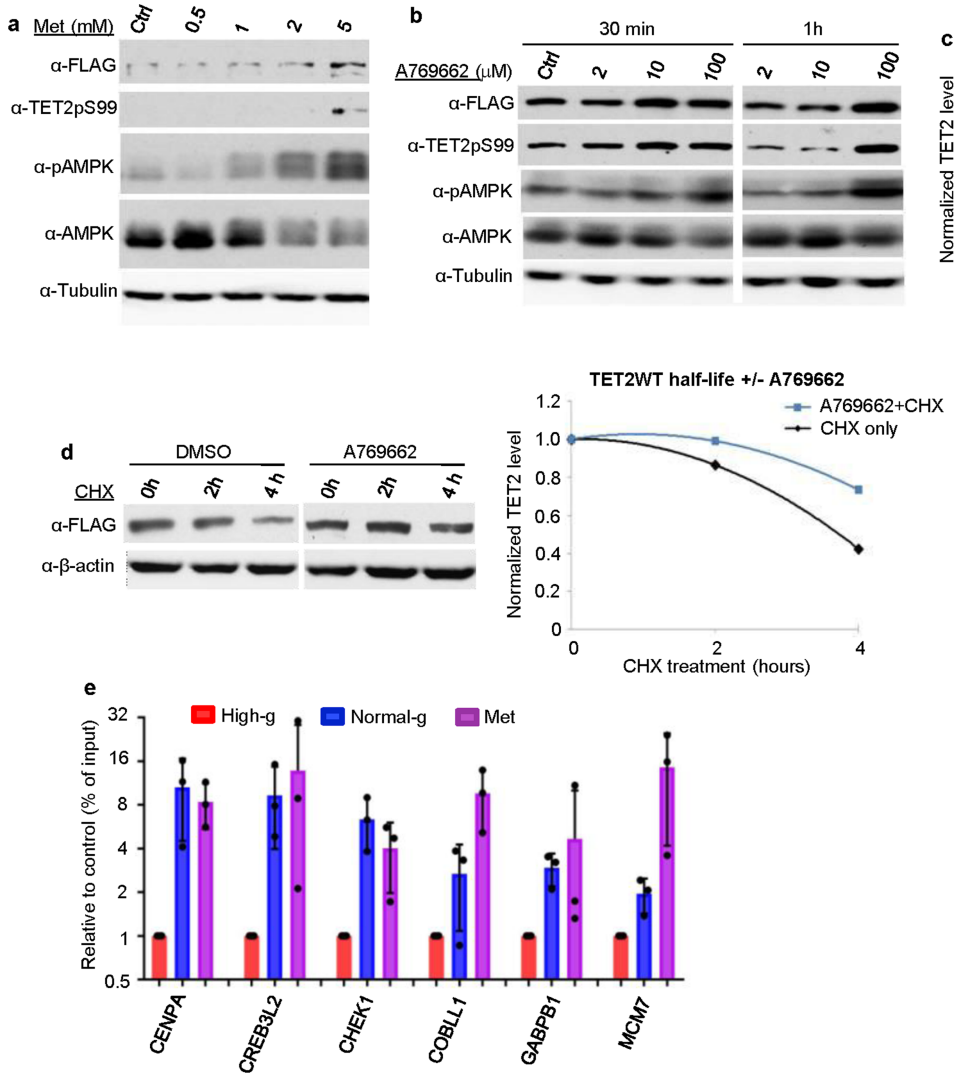
Extended Data Fig. 4 | Purification of full-length TET2 recombinant proteins and determination of TET2 S99 phosphorylation by AMPK. **a**, SDS–PAGE gel showing the quality and quantity of recombinant TET2WT, TET2S99A, TET2SLF and TET2S1205A proteins. **b**, Representation of TET2 S99 phosphorylation levels, as detected by LC–MS/MS after in vitro kinase assay. The peak areas represent the abundance of peptides containing non-phosphorylated S99 (lower lane) or phosphorylated S99 (upper lane). The levels of S99 phosphorylation were much higher after the addition of AMPK in the in vitro kinase assay. **c**, Various amounts of unphosphorylated or phosphorylated TET2S99 peptides were spotted onto nitrocellulose membrane and detected by purified anti-TET2 pS99 antibody (right) or unpurified whole serum (left). The TET2pS99 antibody specifically recognized phosphorylated S99 peptides. Figures in **a**–**c** represent three biologically independent repeats each. **d**, Quantification of the radioactivity of AMPK phosphorylation. At the end of the phosphorylation assay, 20 μg of full-length Flag–TET2 WT and mutant proteins bound to beads were subjected to measurements by Liquid Scintillation Analyzer. The calculation is based on the specific activity of [γ-^{32}P]ATP (800 c.p.m. per pmol) used and the calculated molar amounts of TET2 added to the assay tube. The stoichiometry is around 0.38 mol phosphates per mole of TET2WT. This number decreases markedly if the TET2 protein is mutated (S99A mutant or SLF mutant, both of which disrupt the AMPK-recognizing consensus sequence). TET2WT and the S1205A mutant had comparably high levels of phosphorylation. $n = 3$ biologically independent repeats, with data shown as mean ± s.d.

© 2018 Springer Nature Limited. All rights reserved.



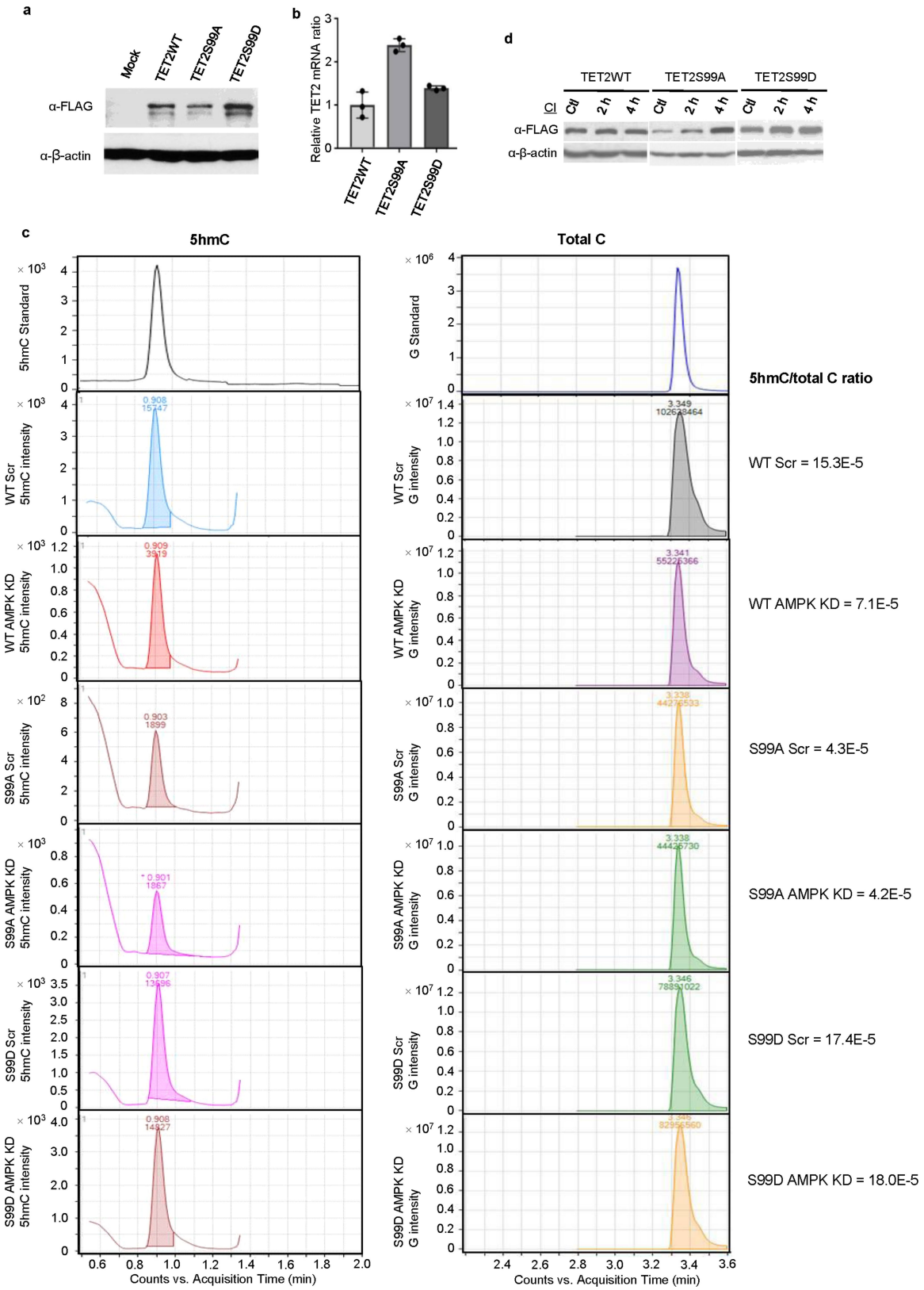
Extended Data Fig. 5 | Increased AMPK activation, TET2 phosphorylation and TET2 stability in cells cultured with normal glucose. a, Representations of TET2 S99 phosphorylation levels detected by LC–MS/MS in A2058-TET2WT cells cultured under high (left) or normal glucose (right). Flag–TET2 from A2058-TET2WT cells cultured under normal or high glucose was purified and then subjected to LC–MS/MS analysis. The peak areas represent the abundance of peptides that contained non-phosphorylated S99 (top) or phosphorylated S99 (bottom). The ratio was calculated as area(phosphorylated S99)/(area(phosphorylated S99) + area(non-phosphorylated S99)), which indicated the level of S99 phosphorylation. S99 phosphorylation (TET2pS99) in cells cultured in normal glucose was significantly higher than the phosphorylation levels observed in cells cultured under high glucose. Data are representative of three biologically independent repeats in each condition. **b**, A2058-TET2WT cells were switched from high to normal glucose for 0, 2 or 4 days. Results showed a steady increase in the levels of pAMPK as well as TET2pS99 and Flag–TET2 during this period. Data are representative of three biologically independent repeats. **c**, Western blot showing increased levels of TET2pS99 and pAMPK in cell lines (PBMC, HUVEC, TF-1) cultured under normal glucose versus high glucose. Data are representative of three biologically independent repeats. **d**, Western blot comparison of TET2, TET2pS99, and pAMPK levels between PBMCs from healthy donors and from patients with diabetes. TET2, TET2pS99 and pAMPK levels were significantly higher in healthy donors than in patients. Data are representative of three western blot repeats.

© 2018 Springer Nature Limited. All rights reserved.



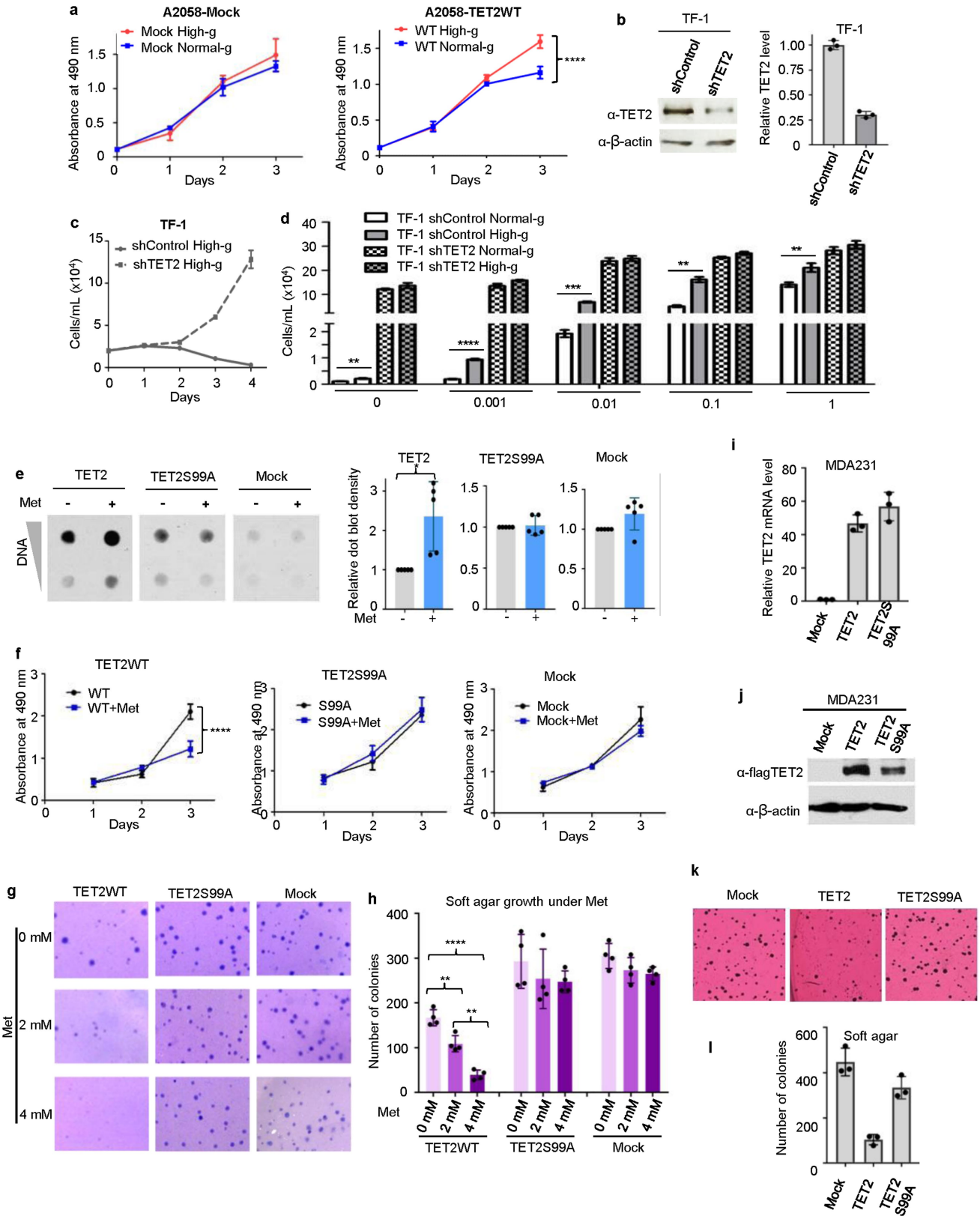
Extended Data Fig. 6 | Effects of AMPK activators (A769662 and metformin) on TET2 protein stability and 5hmC levels. a, A2058-TET2WT cells were treated with metformin and then immunoblotted with the indicated antibodies. Metformin treatment increased the levels of pAMPK, TET2pS99 and Flag–TET2. **b**, A2058-TET2WT cells were treated with 2 μM, 10 μM or 100 μM A769662 (an AMPK activator) for 30 min or 1 h, and then immunoblotted with the indicated antibodies. A769662 treatment increased pAMPK, TET2pS99 and Flag–TET2 levels. **c**, Quantification of normalized Flag–TET2WT half-life in A2058-TET2WT cells treated with CHX and with (+) or without (−) metformin (5 mM) in Fig. 3f. Metformin treatment increased the half-life of TET2. **d**, Left, CHX was used to measure the half-life of Flag–TET2 in A2058-TET2WT cells treated with (+) or without (−) A769662 (100 μM) in high glucose. Right, quantification of normalized Flag–TET2WT treated with CHX and with (+) or without (−) A769662. A769662 increased TET2 stability from 42% to 73% at 4 h after CHX treatment. Data in **a**–**d** are representative of three biologically independent repeats each. **e**, hMeDIP–qPCR showed that metformin treatment increased DhMRs to similar levels as observed in normal glucose in the previously validated genes in Extended Data Fig. 2c. $n = 3$ biologically independent repeats, data shown as mean ± s.d.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 7 | Effects of AMPK knockdown on 5hmC in cells expressing TET2WT, TET2S99A or TET2S99D. a, **b**, Western blot (**a**) and RT–qPCR (**b**) to measure TET2 protein and mRNA levels in cells expressing TET2WT, TET2S99A or TET2S99D. TET2S99A has the lowest protein level (**a**), but it has the highest mRNA level (**b**) amongst the three cell lines. **c**, Representations of HPLC–MS/MS analysis of genomic DNA from A2058 cells expressing TET2WT, TET2S99A or TET2S99D with or without AMPK knockdown (AMPK KD) in Fig. 3i. Left, chromatograms of 5hmC in cells expressing TET2WT or mutants with or without AMPK KD. Right, corresponding chromatograms of G. Total cytosine levels were estimated and equalized to the intensities of G. The ratios (right) were calculated as 5hmC intensity/G intensity, which indicated the cellular 5-hydroxymethylation level. 5hmC levels in A2058 cells expressing TET2WT, but not TET2S99A or TET2S99D, were decreased after AMPK KD. **d**, Comparison of the rescuing effects of calpain inhibitor on TET2WT, TET2S99A and TET2S99D. Data are representative of three biologically independent repeats in **a**, **b**, **d** and three technical repeats in **c**. Data shown as mean ± s.d.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 8 | See next page for caption.

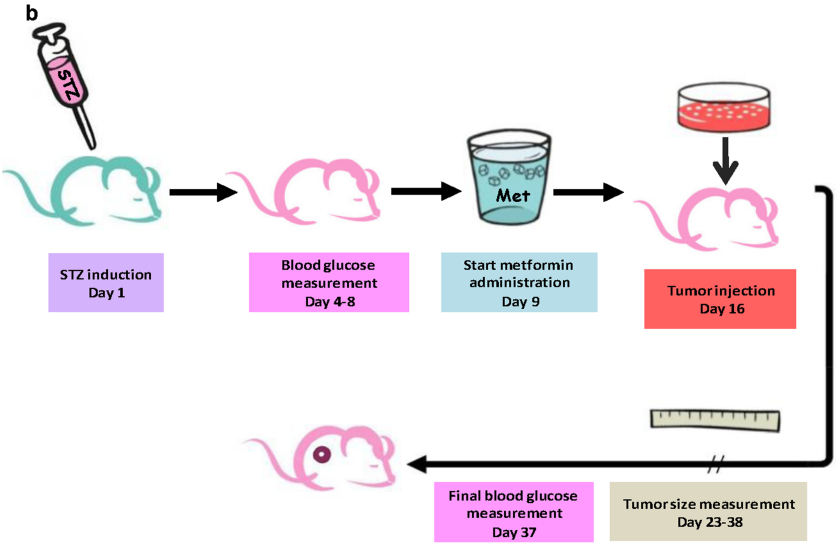
© 2018 Springer Nature Limited. All rights reserved.

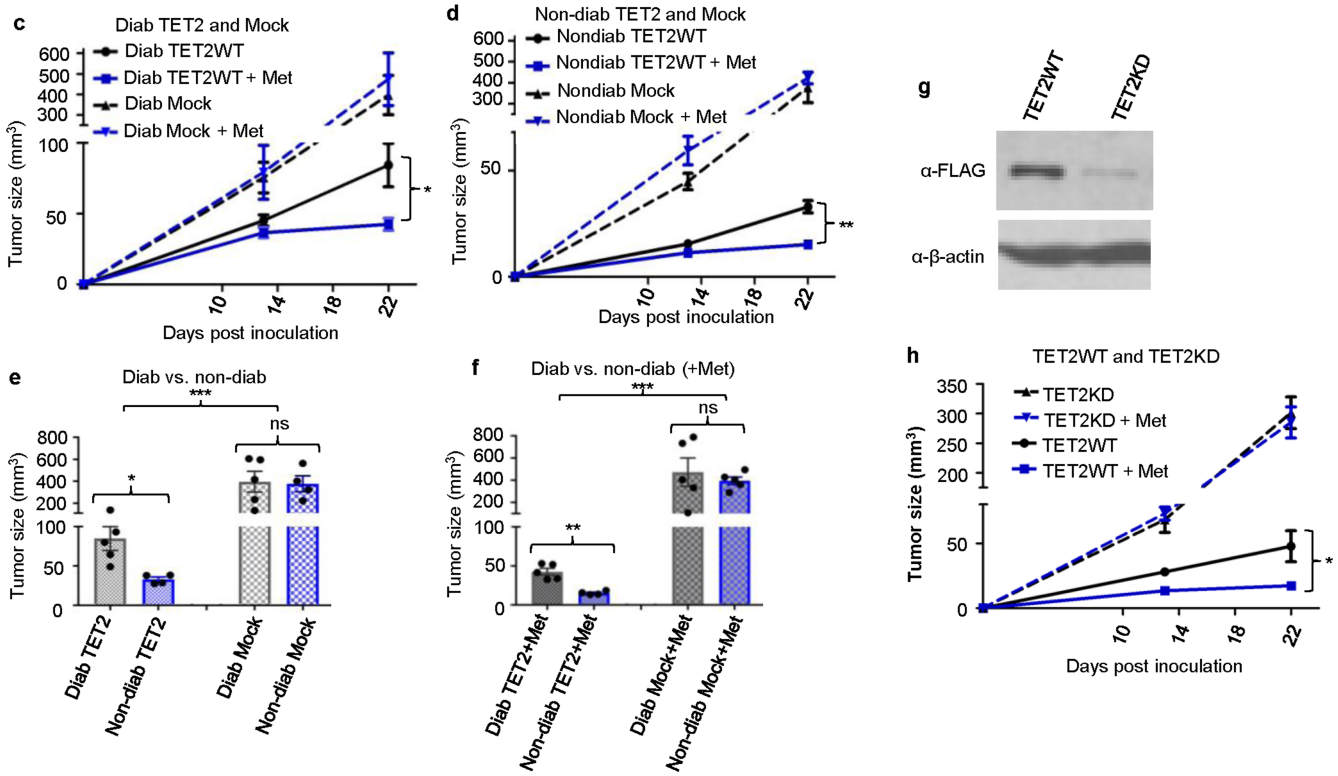
Extended Data Fig. 8 | In vitro characterization of the effects of glucose and the AMPK–TET2 axis on tumour suppression. a. MTS proliferation assay shows that A2058-TET2WT cells grew significantly more slowly under normal glucose than under high glucose, but mock cells showed no difference; $n = 5$ biologically independent repeats, $P = 5.68 \times 10^{-5}$. **b**, TET2 was effectively knocked down (KD) in TF-1 cells, as confirmed by western blot (left) and RT–qPCR (right). **c**, TF-1 TET2 KD cells exhibited cytokine-independent growth under no cytokines, consistent with a previous report[26]. **d**, Proliferation of TF-1 shControl and shTET2 cells treated with high or normal glucose under the indicated cytokine concentrations. TF-1 shControl proliferation was slower under the lower glucose condition, but TF-1 shTET2 cells showed no differential growth rates. *P* values (from left to right) are 0.0078, 1.94×10^{-5}, 3.21×10^{-4}, 0.0011, and 0.0091. **a** and **d** suggest that TET2 is a crucial point through which glucose levels affect tumour growth rates. $n = 3$ biologically independent repeats in **b**–**d**. **e**, Left, dot blot showing 5hmC levels in A2058-TET2WT, A2058-TET2S99A and mock cells with (+) and without (−) metformin (5 mM) treatment. Right, normalization of $n = 5$ biologically independent dot blot repeats; $P = 0.026$. **f**, MTS proliferation assay of A2058 cells expressing TET2WT, TET2S99A or mock treated with (+) or without (−) metformin (5 mM); $n = 5$ biologically independent repeats, $P = 7.18 \times 10^{-5}$. **g**, **h**, Soft agar growth of A2058-TET2WT, A2058-TET2S99A and mock cells treated with 0 mM, 2 mM or 4 mM metformin (**g**) and respective colony counting (**h**). $P = 6.99 \times 10^{-5}$ (top), 0.004 (middle), 0.0013 (bottom). **i**, **j**, mRNA (**i**) and protein (**j**) levels in MDA-MB-231 cells stably expressing mock, TET2WT or TET2S99A. **k**, **l**, Soft agar growth (**k**) and quantification (**l**) of MDA-MB-231 cells stably expressing mock, TET2WT or TET2S99A; $n = 3$ biologically independent repeats. The data suggest that TET2WT suppresses anchorage-independent tumour cell growth in vitro. Prevention of TET2 phosphorylation at S99 results in loss of this TET2-mediated suppression. Two-sided Student's *t*-test, data shown as mean ± s.d. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{****}P < 0.0001$.

© 2018 Springer Nature Limited. All rights reserved.

a

	Non-Diabetic				Diabetic			
Tumor type	TET2	TET2	Mock	Mock	TET2	TET2	Mock	Mock
Metformin	-	+	-	+	-	+	-	+



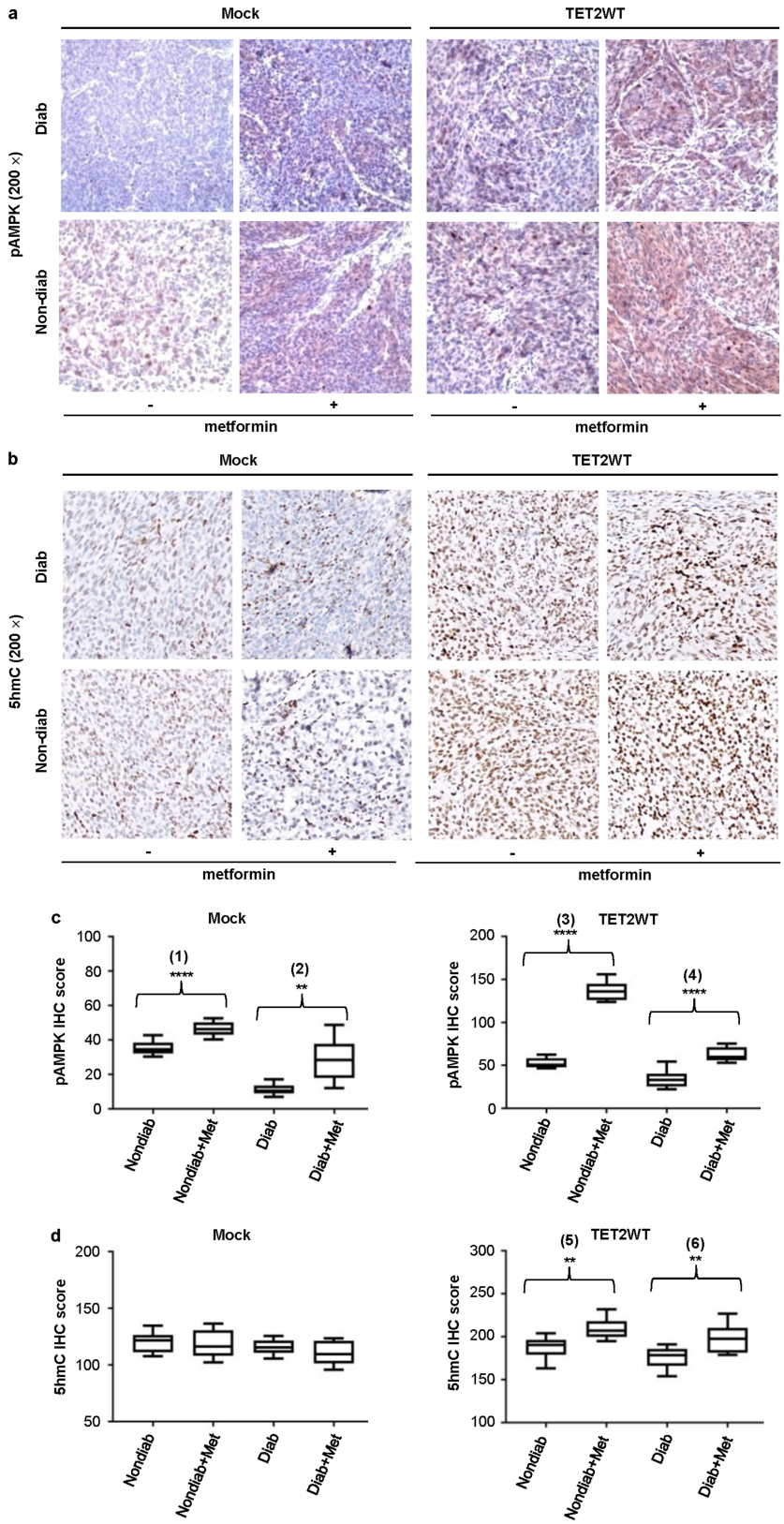


Extended Data Fig. 9 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 9 | The mouse study paradigm and examination of the regulatory function of the tumour-suppressive glucose–AMPK–TET2 axis in vivo. a, Division of mice into eight groups. **b**, Outline of experimental procedure. In brief, nude mice were first induced to develop diabetes, and then transplanted with designated tumour cells. Tumour formation and sizes were documented continuously for three weeks, followed by histological and pathological examination. See more details in Supplementary Information. **c**, **d**, Growth curves of A2058-TET2WT and A2058 mock tumours in diabetic (**c**) and non-diabetic (**d**) nude mice, with and without metformin treatment. $n = 5$, $P = 0.048$ in **c**; $n = 4–5$, $P = 0.0046$ in **d**. **e**, **f**, Comparison between endpoint A2058-TET2WT and mock tumour sizes in diabetic and non-diabetic mice. Mice were treated either with (**f**) or without (**e**) metformin, $n = 4–5$. Tumours from diabetic TET2 groups were significantly larger than that from non-diabetic TET2 groups in both **e** and **f**. However, mock groups showed no difference between diabetic or non-diabetic conditions in either **e** or **f**. $^{*}P = 0.026$ (bottom), $^{***}P = 0.00043$ (top) in **e**; $^{**}P = 0.0023$ (bottom), $^{***}P = 0.00013$ (top) in **f**. **g**, Western blot showed successful TET2 knock down (A2058-TET2KD cells) in comparison with its TET2WT precursor. Data are representative of three biologically independent repeats. **h**, Growth curves of A2058-TET2WT and A2058-TET2KD tumours with and without metformin treatment, $n = 4–5$. The curve indicates that A2058-TET2KD tumours were no longer suppressed by metformin, and grew larger than A2058-TET2WT tumours; $^{*}P = 0.031$. Two-sided Student's *t*-test, data shown as mean ± s.e.m. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < < 0.001$, $^{****}P < 0.0001$; ns, not significant.

© 2018 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | See next page for caption.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Fig. 10 | 5hmC and pAMPK levels in tumour xenografts after metformin treatment in diabetic and non-diabetic mice. a, Representative histology of pAMPK IHC staining in mock and A2058-TET2WT tumours treated with (+) or without (−) metformin (200×); pictures represent four different stained slides in each group. **b**, Representative histology of 5hmC IHC staining in mock and A2058-TET2WT tumours treated with (+) or without (−) metformin (200×); pictures represent four different stained slides in each group. All slides were counterstained with haematoxylin (light blue). **c**, **d**, Statistical analyses of the average score of pAMPK and 5hmC staining in mock and A2058-TET2WT tumours treated with or without metformin. Counts were done on 10 random fields in each group. Metformin treatment increased pAMPK in both mock and A2058-TET2WT tumours xenografted into diabetic and non-diabetic mice; $^{***}P = 6.06 \times 10^{-6}$ (1), $^{**}P = 0.0013$ (2), $^{****}P = 1.04 \times 10^{-11}$ (3), $^{****}P = 3.53 \times 10^{-6}$ (4). However, metformin increased 5hmC only in A2058-TET2WT tumours in diabetic and non-diabetic mice, with no increase in 5hmC in mock tumours under metformin treatment; $^{**}P = 0.0010$ (5), $^{**}P = 0.028$ (6). Box plot: centre lines, median; limits, upper and lower quartiles; top and bottom whiskers, minima and maxima. Two-sided Student's *t*-test, data shown as mean ± s.d. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$.

© 2018 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Correlation data analysis of 5hmC in diabetes, glucose/TET2 co-regulated genes and metformin impact on blood glucose

a

Correlation between 5hmC and various health parameters.

Health Parameters	Beta Coefficient	P-value
Presence of diabetes	-13.51	0.0001
BMI	-0.8	<0.0001
Age	-0.51	<0.0001
Systolic blood pressure	-0.33	0.007
HbA1C	-0.23	<0.05

b

Downregulated in Normal-g vs. High-g	CYB561D1	FDX1	FAM195A	TACC3	NUP85	HSPH1	HAUS7	IMPDH1	ZNF408	SLC25A19	
	GTSE1	RPL13P5	CDC73	PLEKHB2	SLC25A15	FOXM1	ESF1	NUDT19	NOP2	NCAPH	
	LSM10	REXO4	MCM7	NR3C1	PDCD7	C12orf44	TCOF1	WASF1	LYPD1	PAN3	
	MRPL47	MRPL35	HOXB2	RALA	TIPIN	GNG10	BRI3BP	NOP14	RPS27L	TMEM98	
	RINT1	STOM	SOCS7	SASS6	GTF2E2	ITGAE	C1orf109	XPO4	CEP55	FAM73B	
	NME7	ANKRD33B	DOLPP1	ORC4	MRPL46	WDR77	EP400	TMEM201	TEAD4	HHEX	
	TSR1	CXorf40A	NIPA2	EXO1	ZZZ3	ARMC6	PNP	GLRX2	ACOT9	IQCB1	
	SVIP	UNC119	KIF14	GPNMB	HSPA9	TMEM55B	CRADD	QRICH1	MFHAS1	MT1H	
	M6PR	RFC4	SPAG5	CENPA	C15orf61	SNRPA1	GABPB1	MCM6	C19orf48	MAP4K3	
	C5orf30	BCL2L12	KIFC1	LCLAT1	DSCC1	BRIX1	TOMM22	BUB1B	EXOSC8	CRY1	
	NSUN5P2	C10orf32	RAN	ZNF576	MYO19	KIAA0101	ORC6	NCAPD3	HOXC10	C9orf40	
	CHEK1	PIGU	DSN1	FGF13	ACYP1	OPHN1	LSG1	TAF5	HYI	AMOTL2	
	SHROOM2	TCEAL8	MDH1	DARS2	CHST10	C2orf47	STRA13	RNF26	RNASEH2C	EPHA2	
	YRDC	SETD8	WDR36	IPO5	ADM	TRMT11	UBE2C	MRPL15	RNMTL1	GLRX5	
	RWDD3	GADD45A	ANP32B	RAD54L2	SERTAD2	ADO	RFC3	TSNAX	RRS1		
	MTPAP	C16orf80	H2AFX	PPID	METRNL	KIAA0947	ZNF330	C1orf112	MCM3		
	MSTO1	CDC5L	FOXD1	RNF149	RRP7A	PPP6C	FST	CKS1B	SERTAD1		
	VARS2	DDX5	DUS1L	PDRG1	CCT3	LRRC58	DCTPP1	ANAPC4	CHKA		
	EZR	OSBPL3	KIAA1033	C11orf31	CENPN	C1orf9	P2RX5	CRNKL1	SUV39H1		
	TFDP1	DUT	TGS1	CTCF	LIMA1	NSUN2	DDX46	MMGT1	ABL2		
	MAD2L1BP	CDC45	UBE2T	CCNB1	ETF1	EIF3B	HOXB7	SRRM1	HIST1H2BK		
	ASNSD1	UBR7	BTAF1	COIL	FLRT3	CASP7	CDC26	TIPARP	POLA1		
	ZNF783	CKAP2L	PTPLB	PCID2	ZNF721	ZC3HC1	SDC1	EMP1	ERCC6L		
	COX10	ARL4A	PPAT	POGK	DHX33	SRXN1	RBM25	NOL11	POLE2		
	RBM22	C20orf20	TYW3	NUP50	IRS2	ZWINT	PAK1IP1	URM1	ZNF274		
	PNN	NUBP1	DNAJB14	LZIC	TMEM97	PRIM1	CDCA3	BYSL	CDCA5		
	SMC2	SMNDC1	UTP20	SRRM2	WDR45L	BAZ1A	AGPAT9	PIGW	ATAD5		
	MPHOSPH10	MKI67	ADCY6	BTBD6	LIG1	FLNB	CCNT1	MSL1	WDR85		
	CCDC101	ACBD3	SHQ1	DCLRE1A	UQCRFS1	MYBL2	FANCI	BLM	C8orf73		
	TIGD5	NANP	CHERP	NAA40	POP7	CPSF7	POLR3K	PFAS	ATP6V0E2		
Upregulated in Normal-g vs. High-g	ERCC5	SERPINB1	ZCCHC24	KDM5B	PSIP1	GLG1	LGALS1	NEU1	PSAP	RASA1	PPA1
	PAN2	YPEL5	TNC	PCYOX1	IMPACT	NES	TULP3	ETV4	DHX32	DYNLT1	UBAC2
	C3orf32	RABAC1	STX18	RALY	NDUFA9	SPA17	SLK	CXXC5	GTSF1	MAN2B1	DNASE2
	PRKCD	BUD13	MAGED2	PIGH	CCDC104	BRMS1	PLDN	SNX1	GPX4	ARAP1	SF3B14
	SH3BP5	STOML2	FAM60A	WDR45	HEBP1	ATRN	SPSB2	SHISA5	CD97	NGFRAP1	
	DALRD3	EMP3	CALHM2	PKIG	LDLRAP1	AP3B1	SYTL5	GDE1	PLA2G12A	RWDD4	
	B4GALT2	MLKL	APOA1BP	CREBL2	STX16	BAX	RDH11	PIM3	EIF5A	MECR	
	TLCD1	NRIP1	COL16A1	AKR1A1	GNAI2	CTSB	PYCRL	GOLGA5	CCNG1	PFDN1	
	GGPS1	PLOD1	TOR1B	MRC2	COX17	NAPRT1	ZNF512B	PLP1	KDELR3	C17orf48	
	MAML2	SPATA20	CYBA	POLR2G	UQCRQ	BSG	LHFPL2	UPRT	IDE	PFDN5	
	TMEM181	TRIOBP	BLMH	SCO2	TAF6	PREX1	LOC401397	POT1	AMFR	POR	
	SPESP1	C21orf33	G6PD	LOC645212	SNX30	TMEM41B	HAGHL	RALGDS	LEPREL4	LGALS3	
	ELP4	RPS6KB2	KIAA0319L	SLC26A2	TCF7L2	PACSIN3	FAM116A	GTF2A1	SDHA	CDK5	
	GBAS	HOXA5	NFU1	FAF2	LGALS8	TSHZ1	RBL2	GLT25D1	GNAQ	MFSD10	
	LOC147727	NT5E	HAPLN1	ACADVL	ZYG11B	DCP1A	HIBCH	LAMB1	SH3GLB1	VPS11	
	SLC39A1	CEP192	CKB	KCTD3	ARL2	PRKCI	ATG4A	SLC25A6	TTC13	HINT2	
	C10orf125	TPD52L1	C6orf70	SLU7	DAP	CAB39	IGF2BP3	EIF2A	IFI30	VPS4A	
	OCEL1	PAPSS2	S100A3	SCAMP1	NRAS	CETN3	EAPP	FAM108C1	TTC32	ESYT1	
	CYFIP1	PPP1CA	MANSC1	SMYD3	ABR	BCKDK	IL13RA2	ATP5A1	IGBP1	MLL5	
	BBS2	MYH10	ATP6AP1	CLIP4	PMVK	TBCE	SCAND1	SLC44A1	UGGT1	MGAT4B	
	HMGCL	AMOTL1	PRKAR1A	MYBBP1A	MBTPS1	RB1CC1	CD70	TMEM192	FAM82A2	RAB11FIP5	
	FZD7	SNHG8	HSD17B8	HTRA2	TWF2	SV2A	ARL6IP5	DTX3L	PRPSAP2	KIAA1143	
	OSTF1	SIDT2	DECR1	RFXANK	C1orf85	TSPAN31	LOC152217	LCMT1	ATG9A	AP2B1	
	C6orf120	KIAA1191	SNX10	SLC25A46	C5orf54	EIF4B	TTC19	AGGF1	ZMYND8	SDHC	
	FAM125A	C4orf52	POLR3GL	MTFMT	ROBO1	BCKDHA	NDRG3	DHX40	SRP14	MTAP	
	FKBP2	SFRP1	ALKBH6	ARPC1B	EMD	HADH	VPS29	FKBP11	FAM86A	TMEM14C	
	COG4	FAM174A	PTGFRN	ZKSCAN1	MTR	CBR3	NDUFB5	UBE2L6	NECAP1	DHX16	
	MXI1	NFATC2	MPV17	KCNN4	PPP1R7	GABARAP	EDEM1	ACVR1	VPS72	EPS15	
	FAM114A1	SH3BGRL3	PTPN14	RNF181	SERINC1	FADS3	CRELD2	LOC642361	AKR7A2	HOXB3	
	TCTN3	RIN1	C4orf48	DSG2	TSPAN14	COX7A2	RNF123	ALDH2	NT5DC1	BNIP3	

c

Status	Tumor	Metformin	Average Blood Glucose (mM)	Status	Tumor	Metformin	Average Blood Glucose (mM)
Non-diabetic	TET2	-	6.30±0.59	Diabetic	TET2	-	12.94±2.27
	TET2	+	5.65±0.73		TET2	+	14.16±3.76
	Mock	-	5.40±0.84		Mock	-	14.08±3.49
	Mock	+	5.60±0.69		Mock	+	14.76±2.90

a, Univariate regression analysis showing presence of diabetes, BMI, age, systolic blood pressure, and HbA1c in association with 5hmC. These health parameters, particularly the presence of diabetes and HbA1c, are negatively correlated with 5hmC. $n = 28$ healthy donors and 29 diabetic patients. Two-sided t-test.
b, Glucose-modulated and TET2 dependent genes. Five hundred and eighty-five genes that are up- or downregulated in normal glucose compared to high glucose in A2058-TET2WT cells are listed. These genes had differential expression under normal glucose versus high glucose in A2058-TET2WT cells only, and not in A2058-TET2M or mock cells. These genes are defined as glucose-modulated and TET2-dependent. Expression profile of these genes is shown in Fig. 1h.
c, Blood glucose levels in diabetic and non-diabetic mice treated with or without metformin. Blood glucose levels were measured on the last day of experiment. Metformin treatment did not alter blood glucose levels in either diabetic or non-diabetic mice. $n = 4$–5, data shown as mean ± s.d.

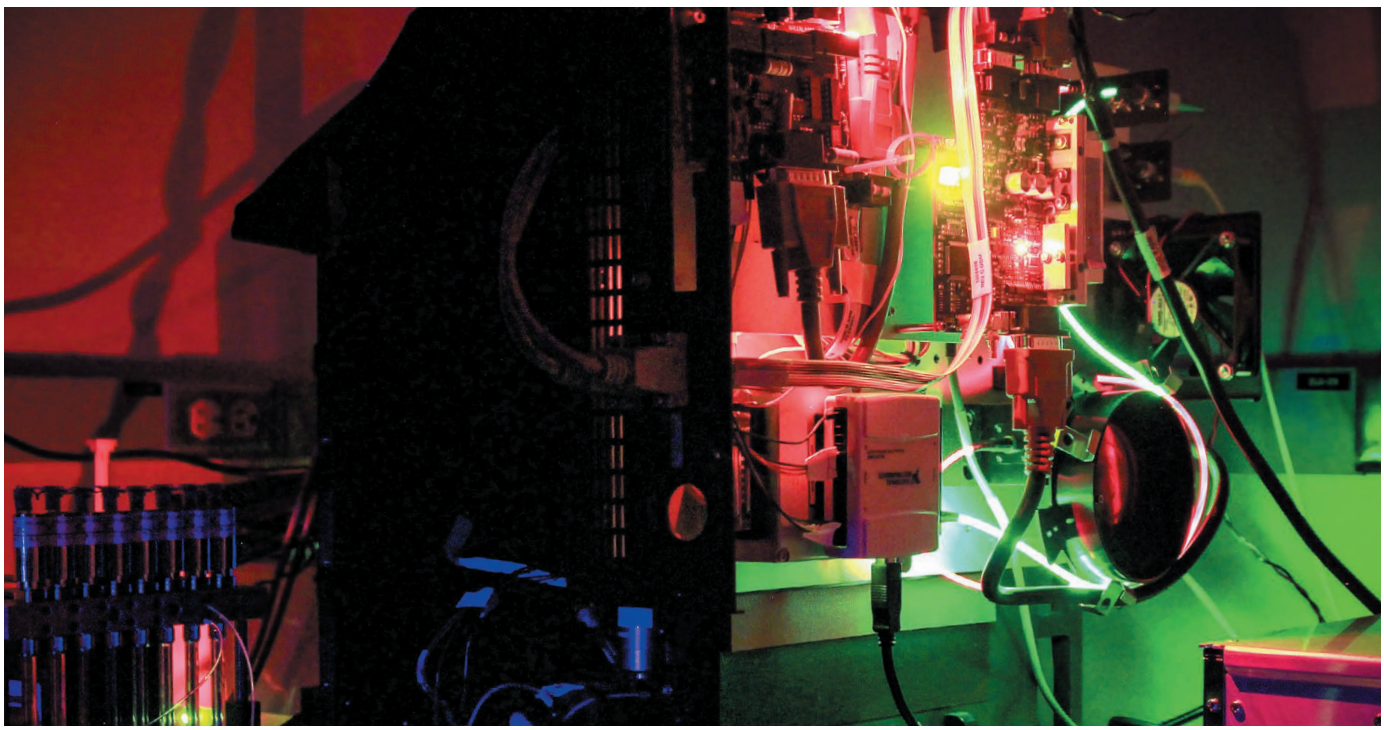
© 2018 Springer Nature Limited. All rights reserved.

TECHNOLOGY FEATURE

HOW TO TEACH AN OLD SEQUENCER NEW TRICKS

Outdated DNA-sequencing machines need not die — researchers can repurpose them to drive next-generation biochemistry studies.

CURTIS LAYTON



This next-generation biochemistry platform, dubbed 'Serenity', was assembled from the components of an old Illumina DNA sequencer.

BY JEFFREY M. PERKEL

In a basement storeroom at Stanford University in California, the guts of a dozen DNA sequencers lie exposed — hundreds of thousands of dollars worth of cameras and lasers, optics and fluid controllers, all scavenged from a late-model, next-generation Illumina DNA sequencer called GAIIx. On the floor, the shell of one old instrument sits empty, picked over like a carcass. "I seem like a hoarder," says Stanford biophysicist William Greenleaf.

But over the past 6 years, this collection has fuelled an effort that has engaged about half of Greenleaf's 18-member lab team. Whereas most researchers use DNA sequencers to, well, sequence DNA, Greenleaf's team is one of a small number that has repurposed the devices for an entirely different goal: to study protein and nucleic-acid biochemistry on a massive scale, from macromolecular interactions and RNA folding to enzyme function.

"It's a revolutionary technology," says Stanford biochemist Dan Herschlag, who uses it to study interactions between RNAs and other molecules. It provides "deep and broad quantitative information", he says, "that allows researchers to build more-precise biophysical and cellular models for molecular interactions, and which is also a critical step towards a truly predictive understanding of biological systems".

Broadly, the work demonstrates what's possible when scientists look into the guts of their hardware — proof that equipment isn't necessarily without value just because it is old or outdated.

But there's a reason such technology development is called bleeding edge: things often go wrong. Sarah Denny, a biophysicist who graduated from Greenleaf's lab this year, chuckles when asked whether her equipment offers 'plug-and-play' simplicity. "Many times when you did an experiment, something would break and you'd have to figure out how to get it to work again," she says. But given the volume of data she could extract, the reward was worth the pain. In Denny's case, her team gained a better understanding of RNA folding. Such is life in the do-it-yourself trenches.

BIOPHYSICS ON A CHIP

When the GAIIx dropped in 2008, it was a hot commodity. Some sequencing centres had dozens of instruments, costing about US$600,000 apiece; in 1 week, a machine could push out 30 billion of the lettered bases that make up DNA. But by 2011, when Greenleaf established his lab at Stanford, the industry had migrated to faster hardware — such as the Illumina HiSeq 2000 — that was more efficient and user friendly, and people were giving their old machines away. "They're basically big paperweights," says Greenleaf. ▶

© 2018 Springer Nature Limited. All rights reserved.

Illumina sequencers automate a sequencing-by-synthesis process. A DNA library is randomly arrayed on a device called a flow cell, and amplified in place to create small clusters of about 1,000 molecules, each representing a single fragment of genetic information. The building blocks of DNA, or nucleotides, are then transferred onto a chip, each containing a unique fluorescent signature and a reversible chemical modification. This process ensures that only a single nucleotide can be added at each cluster. The sequencer then images the array, and on the basis of the colour at each position, 'calls', or reads, the base that was added. The modification is then removed, and the process repeats, allowing the entire sequence to be identified base by base.

THE BASICS OF HACKING

At their core, these instruments are high-end microscopes coupled with liquid handlers that help to move reagents around. Some of their components — particularly cameras, lasers and movable stages — can cost tens of thousands of dollars. In 2009, Christopher Burge, an RNA biologist at the Massachusetts Institute of Technology in Cambridge, realized that it might be possible to repurpose that hardware to do something else.

"The GAII is basically a pumping system that pumps things onto the flow cell and then onto a fancy imaging system," Burge says. "We realized that, well, maybe you could pump other things onto the flow cell."

That, he says, is because the GAIIx was an 'open system', controlled using editable configuration files called recipes and loaded up with reagents that could be changed simply by substituting one tube for another. On the inside, the machine was equally open, with off-the-shelf, third-party components held together with cable ties. "In retrospect, it looks like a high-school science project," says Gary Schroth, a biochemist who directs the genomics-application group at Illumina in San Diego, California, and who collaborated with Burge on his early studies.

New Illumina instruments, by contrast, are more polished, with custom hardware, hardwired control software and barcoded reagents — features that improve the user experience but preclude hacking.

Jacob Tome helped to hack a GAIIx as a graduate student in the lab of molecular biologist John Lis at Cornell University in Ithaca, New York. He recalls "a lot of panicked phone calls" with Illumina representatives, trying to work out the intricacies of the company's instrument-control software. "There's no manual to reprogram the instrument," he says, so working out the logic was largely trial and error.

"One of the most exciting weeks of my PhD was when we just kind of sat there at the sequencer, and we'd write a recipe and then watch the temperature go up and watch what solution it was pumping," Tome says.

Burge, however, had help. Working with Schroth, his team modified the GAIIx recipes to accept new reagents, introduce pauses, adjust the running temperature and alter the imaging parameters. The researchers then applied that modified system to conduct a comprehensive study[1] of the sequencing preferences of a DNA-binding protein in yeast called Gcn4p.

"We showed for the first time that you could turn a sequencer into a biophysical instrument to measure protein–DNA interactions at very high throughput," Burge says.

To do that, Burge's team treated the flow cell — which is usually discarded after sequencing — to regenerate double-stranded DNA at each cluster. They then introduced fluorescently tagged Gcn4p protein to the chip at progressively higher concentrations, and used modified software to trick the sequencer into imaging the flow cell as if it had just completed another round of sequencing chemistry. By downloading and processing the final images, rather than the base calls the instrument typically outputs, the team could measure how much Gcn4p was bound at each place. From that, the group could deduce its affinity for every sequence on the chip — some 440 million measurements in all. "That was a really beautiful, elegant experiment," says Greenleaf. It showed that the GAIIx platform had the potential to provide "all the sorts of measurements I was excited about: kinetic measurements, on-rates and off-rates and equilibrium constants, all the physical measurements that one might need to understand from first principles how DNA–protein interactions work".

Several researchers have used similar hacks to address questions of their own. At the Massachusetts Institute of Technology, for instance, RNA biologist David Bartel looked at messenger RNA (mRNA), the RNA copy of a gene that is used to produce proteins. His team found a link[2] between the number of 'A' bases strung at the end of an mRNA molecule — a feature known as a polyadenylated tail — and how efficiently proteins are produced from genes in animal development.

> ***"When it actually works, boy, it really cooks. Nobody can generate data like that."***

Ilya Finkelstein, a biophysicist at the University of Texas at Austin, hacked a sequencer flow cell to investigate[3] the DNA-binding preferences of a protein called Cascade, used in one version of the gene-editing technique CRISPR–Cas. And some members of Greenleaf's team worked with CRISPR pioneer Jennifer Doudna at the University of California, Berkeley, to study[4] why Cas9 — the most commonly used enzyme in CRISPR — sometimes cuts at the wrong place in the DNA.

Researchers have also worked out methods for transcribing on-chip DNA to create ultra-high-throughput RNA arrays — an approach that could be used to study RNA–protein interactions or to screen libraries of folded RNA molecules called aptamers.

The challenge of making such an array is capturing the transcribed RNA. Lis and Greenleaf independently introduced[4,5] a physical block at the end of double-stranded DNA molecules, stalling RNA polymerase — the enzyme that creates the mRNA molecule — with its dangling transcript. At Weill Cornell Medical College in New York City, RNA biologist Samie Jaffrey's team — led by chemical biologist Nina Svensen, then a postdoc in the lab — used[6] a special, viral RNA polymerase to chemically link the newly synthesized RNA to the chip, and then added a nuclease to degrade the template DNA.

HACKERS AGAINST DISEASE

For Stanford biophysicist Rhiju Das, such RNA arrays have served as platforms for identifying molecules that can diagnose active tuberculosis. The project began when a colleague discovered[7] a signature of three RNAs that can distinguish between dormant and active tuberculosis. The challenge was to pinpoint a single molecule — a 'riboswitch' — that could actually identify which state the disease was in. Such a diagnostic molecule would have to be able to bind all three RNAs, plus a fluorescent reporter, and change shape depending on which ones were present. "It's totally bananas, obviously," Das says.

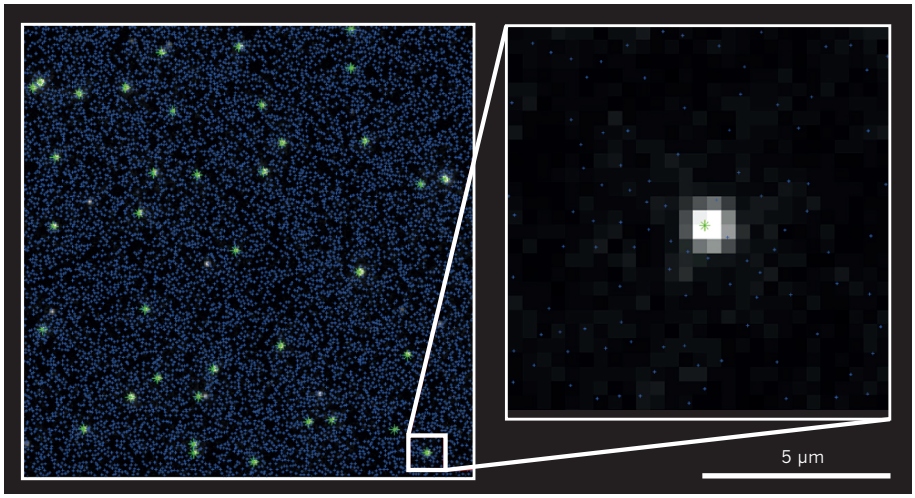
To solve the puzzle, Das recruited the user-base of an online game he co-developed, called eterna. Players are tasked with designing RNAs that can fold in certain ways; the most promising structures get synthesized as part of a DNA-sequencing library, 10,000 at a time, and tested on the GAIIx chip. With 10,000 measurements per design, that's about 100 million data points from each experiment — data that are fed back to the players to hone their skills. Over time, some users have become quite adept, Das says. "These players have now discovered principles that let them get perfect riboswitches, thermodynamically optimal riboswitches for a variety of test cases."

And that includes the OpenTB Challenge. Over 3 rounds of game play, Das's team has tested nearly 27,000 riboswitches from 187 players , identifying several that seem to fit the bill. One, submitted by a retired engineer in Falmouth, Massachusetts, looks like the floor plan of a castle in one state, and like an anchor in the other. Now, with funding from the Bill & Melinda Gates Foundation, Das plans to begin testing some of these designs for use as a pregnancy-test-like diagnostic for tuberculosis.

Meanwhile, a few scientists have taken the next logical step, translating on-chip RNAs into protein. In 2016, Svensen showed[6] she could synthesize short protein sequences called peptides on the array, by exploiting the antibiotic puromycin to tether the peptide to the surface. And in June this year, Greenleaf-lab postdoc Curtis Layton reported[8] developing an assay called Protein display on a Massively Parallel Array to profile 156,140 variants of an enzyme called SNAP-tag. He wanted to learn the subtle rules that relate amino-acid sequences to protein function. "It's been so exciting to see just a

© 2018 Springer Nature Limited. All rights reserved.

CURTIS LAYTON



Fluorescence imaging of a sequencing flow cell reveals binding interactions at specific clusters.

whole new scale of protein functional analysis that we're able to do on these machines," he says.

Exciting, but not easy: Layton spent four or five years developing his method and hardware. Svensen spent two years developing her method, which Jaffrey's lab is now applying to create a whole-human-proteome chip. "To buy a sequencing kit for the GAII back then was $5,000. You can't really make many rounds of optimization if you have to spend that for every experiment you do," she says. As a result, Svensen, now at the Wellcome Centre for Anti-Infectives Research at the University of Dundee, UK, did much of her testing on microscope slides, or by using discarded flow cells from her department's core facility.

A SPACE ODYSSEY

Most GAIIx hackers use the machine largely as is, tweaking the instrument software to make it do what they want. According to Jaffrey, that makes it accessible to anyone with an old GAIIx lying around. "We don't want something only an engineer can use," he says. "Our protocols can be used by any researcher with that machine."

And for its initial RNA work on the GAIIx, the Greenleaf lab was no different. Donated by Stanford geneticist Patrick Brown, the system used in that work, Layton says, was named in honour of the *Star Wars* spacecraft Millennium Falcon — specifically, pilot Han Solo's plea to the ship: "You hear me, baby? Hold together!"

"It's this old machine that we're still trying to run for something it wasn't intended to do," Layton says. "Just don't give us any errors, don't give us any problems. But, when it actually works, boy, it really cooks. Nobody can generate data like that."

Ultimately, the team decided the approach was unsustainable. For one thing, says Layton, unlike modern sequencers with their push-button interfaces, the GAIIx was less Millennium Falcon and more "like flying an Apollo spacecraft". Newer sequencers are also faster, and produce longer reads. Plus, Illumina announced in 2015 that it would be 'end-of-life-ing' the GAIIx, meaning that flow cells, reagents and parts would no longer be available as of July 2017.

The team extracted the system's optics, stage, camera, pumping system and lasers; supplemented the sequencer's built-in illumination system — which lights up the sample at an acute angle from below — with downward-facing widefield illumination; designed a new printed circuit board to operate the lasers; cobbled an autosampler out of GAIIx parts to automate reagent addition; and added an on-board temperature controller.

Most significantly, the team designed new sample holders, so the GAIIx imaging system could accommodate Illumina's new, smaller MiSeq chips. "That gives us the ability to decouple our biophysical measurements of fluorescence and the sequencing," Greenleaf says. (Finkelstein also runs his assays on MiSeq chips, imaging them on a conventional microscope.)

The result is a device to warm an engineer's heart: a compact black cube adorned with what looks like a ripped-open desktop computer on one face. The hardest part, says Layton, was coding the software required to control the hardware, and working out which signals would drive what action. "That was a Herculean effort," he says. But overall, the project appealed to his hobbyist interests in engineering and software design, not to mention biochemistry. Winston Becker, a dual PhD and medical student in the lab, with a master's degree in engineering mechanics, was drawn to the project for similar reasons. "I wanted to do something very physical, very quantitative. And obviously an opportunity to use instrumentation skills and build things was really exciting, too."

But wouldn't it have been easier to just buy a microscope instead? According to Greenleaf, that wasn't really an option — or at least, not an affordable one. "There's really nothing out there," he says, "nothing that will do the sorts of things that we want to do." A custom instrument, he estimates, would probably have cost hundreds of thousands of dollars.

Das, who built his own instrument off Greenleaf's design (and with Greenleaf-lab assistance), says the experience was "so fun". But with no how-to manual, "you have to be willing to go bug people", he says. Still, vexing issues can arise. Das's instrument wouldn't focus properly, for instance. After two weeks of troubleshooting, Das realized he was missing a lens. "It was so dumb," he says. (He has since compiled a 64-page instruction manual. Another researcher, Nick Kaplinsky at Swarthmore College in Pennsylvania, has published detailed blog commentary (see go.nature.com/2l4fdme) describing his work turning a GAIIx into 'RootScope', a microscope for imaging plant roots.)

Sticking with the tradition of naming instruments after fictional starships, Das christened his instrument Red Dwarf, after the British science-fiction series; this team then added a sister ship, Nostromo (*Alien*). Other Greenleaf imagers include Heart of Gold (*Hitchhiker's Guide to the Galaxy*), Borg Cube (*Star Trek: The Next Generation*) and Serenity (*Firefly*). "If you've built an instrument, you will get the glory of the name," Greenleaf says. His lab actually has a dozen sequencers in various states of disassembly throughout the building, most of them acquired for the cost of shipping.

If things go according to plan, that fleet will become ever more integral to Greenleaf's work. His lab is perhaps most famous for developing ATAC-seq[9], a method for assessing how accessible chromatin is to the proteins that bind it, and half the lab is dedicated to that work. Now, Greenleaf hopes to merge the teams, using his hacked sequencers to build mathematical models that accurately reflect macromolecular interactions in the cell itself.

The question is, will other researchers follow suit? Schroth says that between the reagents, the hardware and the analysis tools, these assays are so customized that they could prove difficult to reproduce in other labs, at least without a knowledgeable insider to help.

But there's an easy fix, hackers say: open up modern platforms to exploration. "Then the scientific community could push the instruments in new directions that could benefit everybody," says Burge. Finkelstein says: "There's still a lot of space to play with these sequencers off-label, if you will." ■

Jeffrey M. Perkel *is technology editor for* Nature.

1. Nutiu, R. *et al. Nature Biotechnol.* **29**, 659–664 (2011).
2. Subtelny, A. O., Eichhorn, S. W., Chen, G. R., Sive, H. & Bartel, D. P. *Nature* **508**, 66–71 (2014).
3. Jung, C. *et al. Cell* **170**, 35–47 (2017).
4. Boyle, E. A. *et al. Proc. Natl Acad. Sci. USA* **114**, 5461–5466 (2017).
5. Buenrostro, J. *et al. Nature Biotechnol.* **32**, 562–568 (2014).
6. Svensen, N., Peersen, O. B. & Jaffrey, S. R. *ChemBioChem* **17**, 1628–1635 (2016).
7. Sweeney, T. E., Braviak, L., Tato, C. M. & Khatri, P. *Lancet Respir. Med.* **4**, 213–224 (2016).
8. Layton, C. J., McMahon, P. L. & Greenleaf, W. J. Preprint at bioRxiv https://doi.org/10.1101/342808 (2018).
9. Buenrostro, J. D., Giresi, P. G., Zaba, L.C., Chang H. Y. & Greenleaf, W. J. *Nature Meth.* **10**, 1213–1218 (2013).

© 2018 Springer Nature Limited. All rights reserved.

CAREERS

BLOG Personal stories and careers counsel **blogs.nature.com/naturejobs**

TWITTER Career news, advice and science jobs **https://twitter.com/NatureJobs**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

JEAN CHUNG/GETTY



Water supplies are scarce in parts of the Democratic Republic of the Congo because of poor infrastructure.

LAB LIFE

Science and strife

Five researchers explain how they run productive labs in war-torn or resource-poor regions.

A well-equipped laboratory stocked with reagents and supplied with uninterrupted electricity and unlimited water might seem like a basic requirement for conducting research. But scientists who work in regions that have limited resources or that are riven by conflict cannot take such amenities for granted. They must perpetually seek scarce grants, publish their own journals, form their own scientific societies and — crucially — draw on their deep reserves of resilience. *Nature* asked five such researchers how they run productive labs in the face of electricity shortages, border-checkpoint closures, poor Internet connections and other challenges.

MARLO MENDOZA
Engage with stakeholders

Forestry researcher, University of the Philippines, Los Baños

For the past 13 years, I have been profiling the contamination of the Marilao, Meycauayan and Obando River System (MMORS), which was on the 'Dirty 30' list of the most polluted places in the world in 2007, according to the non-profit organization Pure Earth. There are many polluting industries upstream, including the largest lead smelter in the Philippines, gold smelters, jewellery workshops and tanneries. Downstream are fish farms. We found elevated levels of heavy metals in the water, in the sediments and in fish, especially shellfish, which are sold in the local markets (M. E. T. Mendoza *et al. J. Nat. Stud.* **11**, 1–18; 2012). At least 100,000 people in the municipalities of Marilao, Meycauayan and Obando, and in the metropolitan Manila area, are eating contaminated fish.

There are no toxicologists in the area who can accurately diagnose illnesses connected with heavy-metal ingestion. So when we looked ▶

© 2018 Springer Nature Limited. All rights reserved.

at medical records, there were no entries for heavy-metal poisoning. If we cannot prove that these metals are causing harm to people, it's very difficult to convince policymakers and local executives to take action. We have no local laboratories that can analyse heavy metals found in fish, or in water or blood samples.

Local officials, the governor and some of the mayors were really antagonistic because the fishing industry is a major source of income for these municipalities. I have been very careful, even from the outset, to always update the mayors on our projects, and I am accompanied by local and regional government representatives whenever I do my monitoring activities. I do nothing without their consent and am very transparent in my work.

One of my strategies was to build a network of stakeholders — including national agencies such as the Bureau of Fisheries and Aquatic Resources and the Department of Environment and Natural Resources — that share my concerns. I also built a good rapport with people who live in the region. There are several associations for fishers and leather-makers in these areas, and we work with them and include them in consultations and meetings about water-quality management. Our project helped to have the area declared as a legally designated water-quality management area. That's why we're able to continue our work.

We used funding from Pure Earth to do regular longitudinal sampling in sections of the river system, including of sediment, water, fish and other aquatic life. There's a problem collecting data and samples, because it is costly and the national and local governments have limited funds. There is also no single repository of data with which monitoring can be more effectively planned and analysed.

Our monitoring results were included in a Pure Earth database that was shared with other stakeholders, including regional environmental-management offices and local government units. In turn, this encouraged those agencies to conduct studies to complement our work and to share their data. So I was able to get money from the Asian Development Bank, Green Cross Switzerland and the Hong Kong Shanghai Banking Corporation, as well as a small amount from the Coca-Cola Company, to conduct environmental monitoring — including assessment of heavy metals in selected aquatic organisms.

EMMANUEL I. UNUABONAH
Use available resources

Materials chemist, Redeemer's University, Osun State, Nigeria

Potable water is a challenge for us here in Africa and across the world: around 1.8 billion people worldwide get their drinking water from a source that is polluted with faeces. As part of our work, we are developing hybrid clay composites to adsorb enteric bacteria, such as *Escherichia coli*, *Salmonella* species and *Vibrio cholerae*, from water. We also use composites made from readily available materials such as kaolinite clay, papaya seeds and plantain peels to extract heavy metals from water.

We are not funded by the government. On average, for close to 100 days a year, we have no electricity. We have an alternative utility on campus, so when the power goes off at the national grid during work hours, the generator comes on. If we get lucky with timings, we are guaranteed 36 hours of uninterrupted power to run experiments. But when the generator isn't running and the grid power has gone off, we just have to wait.

Sometimes I use my salary to fund my research and to keep our students. Then I have to struggle to write international grants. I'm so grateful to The World Academy of Sciences in Trieste, Italy; the last grant it gave in 2014 (for US$63,230) took care of stipends, school fees and research expenses for the students, and we used part of it to buy equipment. A colleague at the University of Edinburgh, UK, sent us a $600 bacteria-testing kit last year, but we can't use it now because a related microscope part was damaged by a power surge.

We have a lot of wonderful ideas, wonderful theses just hanging about the shelves, but nobody's utilizing them. Some young scientists developed cheap electrical power systems from electronic waste materials, but they don't have the money to develop them further. Nigeria has a thriving oil industry, but the government's Petroleum Trust Development Fund uses oil-industry proceeds mostly to fund scholarships for Nigerian students abroad, and spends very little on scientific research.

KALULU MUZELE TABA
Aim for the possible

Organic chemist, University of Kinshasa, Democratic Republic of the Congo

Our research seeks to solve problems that have societal impacts, such as malaria, which is endemic in Kinshasa, the capital of the Democratic Republic of the Congo (DRC). People in the poorest areas of Kinshasa are growing about 55 different plants, including citronella and papaya, to try to treat the symptoms of the disease. We thought, why not investigate these plants? We tested eight of the most-used plants and showed that extracts and metabolites had considerable antimalarial activities. We have a small booklet in French and in the local language, Lingala, that we send to people to explain how to use these plants more effectively. We are also studying plant extracts that can be used to treat antibiotic-resistant *Mycobacterium tuberculosis* (K. B. M. Jose *et al. Med. Clin. Rev.* **4**, 5; 2018), using a grant from The World Academy of Sciences.

We don't get money from the state. In our lab, we don't have equipment. We don't have money to buy solvents. Water is available between 5 a.m. and 7 a.m., so we have a container that collects water at night, and during the day we have a pump. For electricity it's much harder. In the middle of the day, it can come and go many times, and you hope it won't go while you're working. We used to have a small generator as a backup for computers, but it's broken.

"To scientists working in comparable circumstances, I would say that although it's hard, it's not impossible."

We don't fold our hands and cry and say that things will get better. We do whatever we can. Most of the time we buy our own reagents and solvents with our salaries. We try to motivate our master's and PhD students by finding a way to collaborate with the outside, writing to foreign labs to see whether our students can get overseas fellowships. One student, Joséphine Ntumba, went abroad three times, to the Catholic University of Louvain in Louvain-la-Neuve, Belgium. She has completed her PhD and teaches at the University of Kinshasa. I did my PhD at Northwestern University in Evanston, Illinois, and then went to the Max Planck Institute for Coal Research in Mülheim an der Ruhr, Germany. It was hard to come back. It was not only material, but mental too. I knew that some things would be impossible, but I feel that I have to contribute and inspire young people in science.

For the past five or six years, I have been the editor-in-chief of the journal *Congo Sciences*, which I co-founded. We started it because we wanted to bring visibility to research done in the country. The journal was financed for some time by the Academy of Research and Higher Education (ARES) in Brussels. The academy has stopped funding it now, but we are still publishing the journal.

For the past ten years, I've been trying to create an academy of sciences for the DRC, similar to the American Association for the Advancement of Science in Washington DC. We have to try to get scientists together and to speak as one voice, and then perhaps the state can start understanding that financing research at the university is important. These are some of the things that make me feel happy that I came back home. Maybe I lost a lot as a scientist, but as a Congolese, I hope I can do something for my place, and for the world.

To scientists working in comparable circumstances, I would say that although it's hard, it's not impossible. Know that you should find maybe not the best solution, but the least-worst one.

© 2018 Springer Nature Limited. All rights reserved.



GGateway students on the training scheme in information technology funded by the Basque government.

MOHAMMED SAFIA

RASHA ABU-SAFIEH

Choose the positive

Computer engineer and co-founder of GGateway, Gaza Strip

I co-founded GGateway, a social-enterprise company in the Gaza Strip that provides outsourcing services for information and communications technology (ICT) around the world. We offer training and employment to recent university graduates in Gaza with ICT degrees. Our main goal with GGateway is to help people to have a source of income. With the shortage of clean water, the polluted sea and the blockade imposed by Israel and Egypt since 2007, living conditions here are dire.

We came up with the idea of GGateway in 2012. In November 2013 we launched a pilot, and in February 2015 we got the green light from the Korea International Cooperation Agency for a $1.3-million grant to fund our plan. That was one of the biggest, happiest things that ever happened to us.

We were running a pilot project for the United Nations Relief and Works Agency for Palestine Refugees in the Near East (UNRWA), but in July and August 2014 we had the 50 days of conflict with Israel. That was difficult: the bombing went on all day, there was no electricity and we had limited access to water and food. There was no safe zone, no safe area. So we had to stop our first project.

Two days after a ceasefire ended the conflict, we wrote up another eight new concepts for different projects, and the UNRWA agreed to operate three. We could either be positive or cry all day. We chose the positive side and to move on.

The cables that we use for all our technical networks are on the list of items that Israel does not allow to enter Gaza. With the support of the UNRWA, it took us nearly four months to bring them in from Israel. Without the UNRWA, it would have taken us a year or more. We also use a generator because most days we get electricity for two to four hours.

The overall unemployment rate among graduates in Gaza with ICT degrees is 70%. Among female ICT graduates, it's 92%. We applied for grants that focus on vulnerable women, and got one from the Basque government in Spain to train 60 female graduates for jobs. We also got a grant from the US-Middle East Partnership Initiative to empower and train 300 ICT graduates to become professional freelancers, and won a $3-million grant in June from the World Bank to train students and software engineers. We have contracts with UN Women and UN Habitat.

We are able to travel outside Gaza only two to three times a year, maximum, and sometimes not at all; we need to apply for an exit permit from Israel with support of the UNRWA. Sometimes we are accepted, sometimes rejected; no reason is given. We often lose opportunities if we are registered for a conference, for example, or for training.

When we see what we are doing — that it's changing people's lives, despite all of the difficulties around us — it makes us feel good.

ELIZABETH TILLEY

Focus on small but crucial changes

Sanitation economics researcher, University of Malawi, The Polytechnic, Blantyre

I came to Malawi in 2015 after 9 years as a project officer and PhD student at the Swiss Federal Institute of Aquatic Science and Technology in Dübendorf. I had worked on sanitation projects in Nepal, South Africa, Tanzania and Nicaragua. Most of my work in Malawi now is teaching and supervising master's and PhD students on such projects as making fuel briquettes out of dried faecal sludge. We work on 'shit-flow diagrams' — trying to map and understand where excreta is being generated and how much of it is being treated.

We have a very bad Internet connection, and it's a barrier to downloading files or making Skype calls. We don't have subscriptions to journals. We have 30 computers for 4,000 students. Paper and photocopying are very expensive. We go days without water to even flush the toilets. At the university, we don't have toilet paper, so I bring my own each day and I keep a secret bottle of soap.

The research agenda in Malawi is driven by big donors from the global north, including national governments such as Norway and Japan, and private donors and non-governmental organizations. Very little funding goes to African researchers for work on topics that they've identified themselves. The fact that northern countries offer funding opportunities to those in the global south is an excellent form of development.

But some proposals call for the participation of a southern partner with no requirements for the division of funding. This means that the southern partner is sometimes given a limited budget for limited work that has limited impact.

I would encourage northern researchers to think about doing sabbaticals in African universities. It gives the southern researcher a chance to focus on publications or research, to be exposed to new ideas and methods, and to connect with a broader network, and the northern researcher can learn how things operate in the south and appreciate what works well at home.

When you start to think about how crushing the whole system is, you can go crazy. I had a student who just wrote to tell me that he got into a master's programme in the United States, and to thank me for the reference letter. He's so excited, and that's the kind of thing I can hold on to for a couple of months. ■

INTERVIEWS BY JOSIE GLAUSIUSZ

These interviews have been edited for clarity and length.

© 2018 Springer Nature Limited. All rights reserved.

SURROGATE

Total recall.

BY GRIFFIN AYAZ TYREE

My body was at the dining table where I'd left it.

Postured differently now: bolt upright, one hand clenched around the stem of a wine glass with the fingers going numb — it was a wonder I hadn't snapped the thing in half.

It hadn't been this tense before.

Nick squirmed in his seat, started to say something but then thought better of it. He was a paying client; no need to explain the aborted session. I pushed back my chair and stood.

"Can … can you tell her I'm sorry?" he stammered.

I felt it then, the knot in my stomach, the lingering heat on my cheeks where I'd flushed in anger.

"Argument?" I asked.

A nod.

"It's common. There's a lot to rehash the first time."

I ejected his wife's loop-drive from my cranial rig and laid it on the table.

"Same time next week?" His voice was hopeful and a little strained.

I smiled. Of course. How else would I pay the bills?

On the train home, a sudden craving for noodle soup. Shoyu broth, sliced pork, soft-boiled egg — nothing I'd ever liked.

Is that you, Lana?

Only the steady hum of the train in reply. I leant my head against the window and dozed off while rain battered the plexiglass. When I came to, the weather had cleared and Lana was ready to talk.

What am I?

She knew the answer to that; everyone did. Neural connections from a dying brain, stored in binary and back-translated into another body. Next question.

Why am I here?

Because running someone's persona through a cranial rig leaves traces behind — electric patterns worn like footprints in snow.

Huh.

She was taking it well, relatively. Most would have filled my blood with stress hormones by now.

My stomach growled and my thoughts — *our* thoughts, I guess — drifted back to food. Lana loved ramen but Nick was never in the mood, wouldn't even humour her. When she got sick he'd bring it home everyday, in big Styrofoam containers, but by then the chemo had her doubled-over and throwing up most of the time.



So what happens now?

Right. The question I'd hoped to put off as long as I could. At least I could misdirect.

Now we eat.

Tyler never asked existential questions. He was six when he died; freak accident, drowned in a pool. Never wondered what it meant to be 'reset' with the cranial rig. Too young to know that was standard procedure.

He was a brat, though: he'd throw a fit every time he saw a toy (dinosaurs were his favourite) and he wouldn't let up unless I bought it for him. There was a toy store across from the ramen bar; I tried not to look at it for too long.

That's classic Nick, Lana said between big gulps of savoury broth, as if it was her doing the eating and not me. *I'm not even cold in the ground and he's out looking for someone younger.*

Tyler's parents would hire me for six hours at a time, whole days spent at the park or the zoo or the state fair. When the session was over I'd find myself holding a new action figure or oversized soft toy — the kid was true to form.

He was a brat, like I said. He was tiring.

And once the parents realized they had a choice — when to meet, *whether* to meet — they called less often. Eventually, they stopped altogether.

As gently, as inconspicuously as I could, I slid my left hand up my neck to feel for the small oval button on the side of the cranial rig …

He expects me to be fine. Like we just 'separated'. Like I'm not even dead.

It was the perfect arrangement for the bereaved, really: to them, their son was still alive on a loop drive somewhere — nothing lost, nothing to mourn. And they could see him again whenever they wanted.

To the other party, though, it was … Lana put it best: *The worst parts of being alive, none of the perks of being dead. It sucks.*

I was sweating then — I think it was the soup, the heat. I don't think it was nerves.

Then why were my hands shaking?

Look, whatever you have to do, just do it quickly.

It always came to this: finger over the button, poised to erase an afterthought of someone else's refusal to grieve.

It always came to this, and I always made the same choice.

We bought a housekeeping unit when I was a kid, one of those low-end models with a hacked-together AI. It had a real problem with steps; it would fall to the side and give off a strange synthetic groan.

Naturally I sent it up and down stairs every chance I got.

My mother would chide me for it: "Don't be cruel, child."

But it isn't real, I'd say, it wouldn't even matter if I knocked it over.

"It would matter a great deal to *you*. To the person you are."

Mom died without a loop drive. *Gone and lost forever*, some would say. But there are parts of her I'll carry to my grave.

After dinner I emerged into the cold air and made a bee-line for the toy shop. Lana helped me pick out a stuffed dinosaur from the display, a little blue Stegosaurus.

He'll love it.

Yeah, if he hasn't already seen.

It was a ten-minute walk to the train station through empty streets. I used to hate that kind of thing, the solitude, the fear — but I haven't felt alone in a very long time. ■

Griffin Ayaz Tyree *is a psychiatry-bound medical student. He lives in San Diego with his partner, and several small birds who leave droppings on their patio. He tweets as @ayazyouyaz.*

ILLUSTRATION BY JACEY

NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm

© 2018 Springer Nature Limited. All rights reserved.